SCIENTIFIC AMERICAN

DISPLAY UNTIL JANUARY 22, 1998

$4.95

SPECIAL ISSUE

THE solid-state CENTURY

the past, present and future of the transistor

MICROCHIPS AT THE LIMIT: HOW SMALL? HOW FAST?

RISE OF THE DUMB PC AND THE SMART PHONE

IGBTs: LOGIC MEETS MUSCLE

QUANTUM COMPUTING

ATOM-SCALE ELECTRONICS

100 years

Copyright 1997 Scientific American, Inc.

SCIENTIFIC AMERICAN

SPECIAL ISSUE/1997
THE SOLID-STATE CENTURY

FIFTY YEARS OF HEROES AND EPIPHANIES 7
GLENN ZORPETTE
Introducing an epic of raw technology and human triumph.

1 THE TRANSISTOR

BIRTH OF AN ERA 10
MICHAEL RIORDAN AND LILLIAN HODDESON
When three Bell Labs researchers invented a replacement for the vacuum tube, the world took little notice—at first. An excerpt from the book *Crystal Fire*.

THE TRANSISTOR 18
FRANK H. ROCKETT
From *Scientific American,* September 1948: this early detailed report on the significance of the transistor noted that "it may open up entirely new applications for electronics."

COMPUTERS FROM TRANSISTORS 24
Inside every modern computer or other data-processing wonder is a microprocessor bearing millions of transistors sculpted from silicon by chemicals and light.

DIMINISHING DIMENSIONS 24C
ELIZABETH CORCORAN AND GLENN ZORPETTE
By controlling precisely how individual electrons and photons move through materials, investigators can produce new generations of optoelectronic gadgets with breathtaking abilities.

HOW THE SUPER-TRANSISTOR WORKS 34
B. JAYANT BALIGA
Think of it as a transistor on steroids. Insulated gate bipolar transistors can handle enough juice to control the motors of kitchen blenders, Japan's famous bullet trains, and countless items in between.

WHERE TUBES RULE 44
MICHAEL J. RIEZENMAN
Surprisingly, transistors have not made vacuum tubes obsolete in all applications. Here's a look at the jobs that only tubes can do.

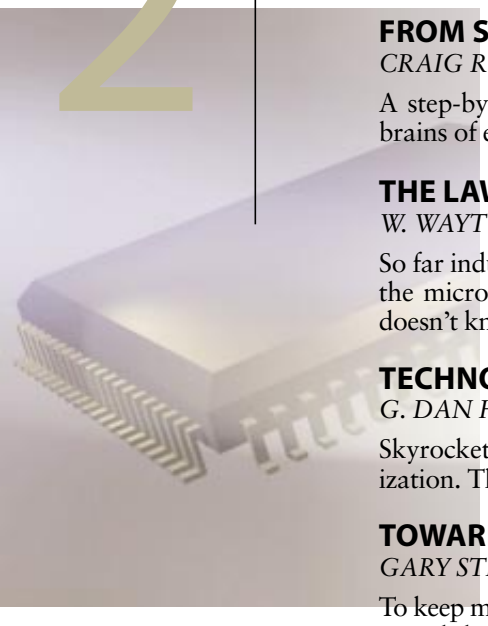
THE FUTURE OF THE TRANSISTOR 46
ROBERT W. KEYES
For the past 50 years, transistors have grown ever smaller and less expensive. But how low can they go? Are there barriers to how much more these devices can shrink before basic physics gets in the way?

Cover and Table of Contents illustrations by Tom Draper

Scientific American The Solid-State Century (ISSN 1048-0943), Special Issue Volume 8, Number 1, 1997, published by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 1997 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. To purchase additional quantities: 1 to 9 copies: U.S. $4.95 each plus $2.00 per copy for postage and handling (outside U.S. $5.00 P & H); 10 to 49 copies: $4.45 each, postpaid; 50 copies or more: $3.95 each, postpaid. Send payment to Scientific American, Dept. SSC, 415 Madison Avenue, New York, N.Y. 10017-1111. Canadian BN No. 127387652RT; QST No. Q1015332537.

Copyright 1997 Scientific American, Inc.

2 INTEGRATION: THE TRANSISTOR MEETS MASS PRODUCTION



FROM SAND TO SILICON: MANUFACTURING AN INTEGRATED CIRCUIT 56
CRAIG R. BARRETT

A step-by-step guide through the machines and processes that turn silicon wafers into the brains of electronic devices.

THE LAW OF MORE 62
W. WAYT GIBBS

So far industry has kept pace with the 30-year-old observation by Gordon E. Moore, father of the microprocessor, that the density of integrated circuits grew geometrically. But even he doesn't know how much longer that can last.

TECHNOLOGY AND ECONOMICS IN THE SEMICONDUCTOR INDUSTRY 66
G. DAN HUTCHESON AND JERRY D. HUTCHESON

Skyrocketing development and manufacturing costs might eventually curb further miniaturization. The good news is that computing power and economic growth could still continue.

TOWARD "POINT ONE" 74
GARY STIX

To keep making devices more compact, chipmakers may soon have to switch to new lithographic tools based on x-rays or other technologies. Progress, however, can be slow and expensive.

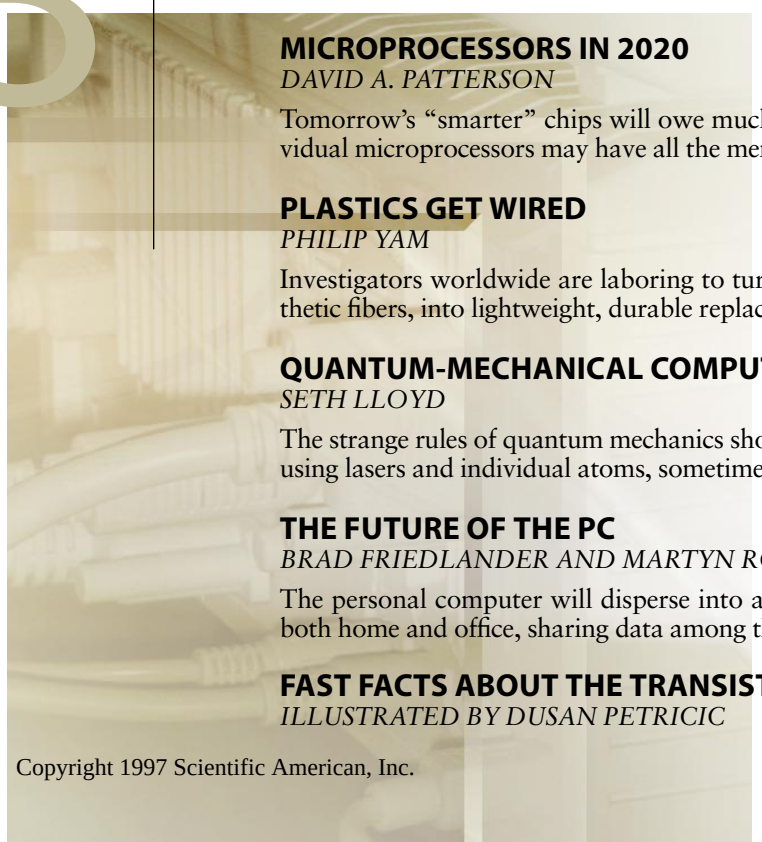
TACKLING TYRANNY 80
ALAN GOLDSTEIN

"The tyranny of numbers" described the showstopping problem of linking millions of microcomponents into a working machine. Then Jack Kilby hit on the idea of the integrated circuit.

THE SEMICONDUCTING MENAGERIE 82
IVAN AMATO

Silicon may be king of the chips, but there are pretenders to the throne. Gallium arsenide and other semiconductors have their uses, particularly for emitting light.

3 THE REVOLUTION CONTINUES



MICROPROCESSORS IN 2020 86
DAVID A. PATTERSON

Tomorrow's "smarter" chips will owe much to the smarter design of their architecture. Individual microprocessors may have all the memory and power of full computers.

PLASTICS GET WIRED 90
PHILIP YAM

Investigators worldwide are laboring to turn organic polymers, the stuff of plastics and synthetic fibers, into lightweight, durable replacements for silicon and metals in circuits.

QUANTUM-MECHANICAL COMPUTERS 98
SETH LLOYD

The strange rules of quantum mechanics should make it possible to perform logical operations using lasers and individual atoms, sometimes at unrivaled speeds.

THE FUTURE OF THE PC 106
BRAD FRIEDLANDER AND MARTYN ROETTER

The personal computer will disperse into a personal network of savvy, doting appliances at both home and office, sharing data among themselves and, cautiously, with others.

FAST FACTS ABOUT THE TRANSISTOR 112
ILLUSTRATED BY DUSAN PETRICIC

Copyright 1997 Scientific American, Inc.

Getting Small but Thinking Big

Proving the adage that great things come in small packages, transistors have grown only more important as they have shrunk. At the clunky stage of their early development, they seemed like mere alternatives to vacuum tubes. Even so, they led inventors to design more compact versions of radios and other conventional gadgets. When transistors could be integrated by the thousands and millions into circuits on microprocessors, engineers became more ambitious. They realized that they could mass-produce in miniature the exotic, room-filling machines called computers.

With every step down in transistor size, technologists found inspiration and capability to build microelectronic devices for jobs that were not only once impossible but inconceivable. Today transistors and other solid-state devices live inside telephones, automobiles, kitchen appliances, clothing, jewelry, toys and medical implants. This is the Information Age not only because data processing is so common but because it is increasingly possible to cast all problems as matters of data manipulation—to see the world as a frenzy of bits waiting to be tamed.

Three decades ago John Updike read an issue of *Scientific American* on materials and wrote several verses, including this one:

> The Solid State, however, kept its grains
> Of Microstructure coarsely veiled until
> *X-ray diffraction* pierced the Crystal Planes
> That roofed the giddy Dance, the taut Quadrille
> Where Silicon and Carbon Atoms will
> Link Valencies, four-figured, hand in hand
> With common Ions and Rare Earths to fill
> The lattices of Matter, Glass or Sand,
> With tiny Excitations, quantitatively grand.
>
> —from "The Dance of the Solids," by John Updike (collected in *Midpoint and Other Poems*, Alfred A. Knopf, 1969)

I hope readers of this special issue will find in it something at which they too can wonder.

JOHN RENNIE, *Editor in Chief*
editors@sciam.com

A NOTE ON THE CONTENTS

Some of the articles in this issue previously appeared in a different form in *Scientific American:* "Diminishing Dimensions," "The Future of the Transistor," "Technology and Economics in the Semiconductor Industry," "Toward 'Point One,'" "Microprocessors in 2020," "Plastics Get Wired" and "Quantum-Mechanical Computers."

The original authors and the editors have updated or thoroughly rewritten those articles to ensure that today's readers are receiving the most current information on the subjects. *—The Editors*

SCIENTIFIC AMERICAN®
Established 1845

Scientific American *The Solid-State Century* is published by the staff of Scientific American, with project management by:

John Rennie, EDITOR IN CHIEF

Michelle Press, MANAGING EDITOR

Glenn Zorpette, PROJECT EDITOR
Sasha Nemecek, ASSISTANT EDITOR

STAFF WRITERS: W. Wayt Gibbs; Gary Stix; Philip Yam

Art
Jessie Nathans, ART DIRECTOR
Adrienne Weiss, ASSISTANT ART DIRECTOR
Lisa Burnett, PRODUCTION EDITOR
Bridget Gerety, PHOTOGRAPHY EDITOR

Copy
Maria-Christina Keller, COPY CHIEF
Molly K. Frances; Daniel C. Schlenoff;
Terrance Dolan; Katherine Wong;
William Stahl; Stephanie J. Arthur

Administration
Rob Gaines, EDITORIAL ADMINISTRATOR
Sonja Rosenzweig

Production
Richard Sasso, ASSOCIATE PUBLISHER/
VICE PRESIDENT, PRODUCTION
William Sherman, DIRECTOR, PRODUCTION
Janet Cermak, MANUFACTURING MANAGER
Tanya DeSilva, PREPRESS MANAGER
Silvia Di Placido, QUALITY CONTROL MANAGER
Madelyn Keyes, SYSTEMS MANAGER
Carl Cherebin, AD TRAFFIC; Norma Jones;
Kelly Mercado

Circulation
Lorraine Leib Terlecki, ASSOCIATE PUBLISHER/
CIRCULATION DIRECTOR
Katherine Robold, CIRCULATION MANAGER
Joanne Guralnick, CIRCULATION PROMOTION MANAGER
Rosa Davis, FULFILLMENT MANAGER

Advertising
Kate Dobson, ASSOCIATE PUBLISHER/ADVERTISING DIRECTOR
OFFICES: NEW YORK:
Meryle Lowenthal, NEW YORK ADVERTISING MANAGER;
Kevin Gentzel; Thomas Potratz; Timothy Whiting.
DETROIT, CHICAGO: 3000 Town Center, Suite 1435,
Southfield, MI 48075; Edward A. Bartley, DETROIT MANAGER;
Randy James. **WEST COAST:** 1554 S. Sepulveda Blvd.,
Suite 212, Los Angeles, CA 90025;
Lisa K. Carden, WEST COAST MANAGER; Debra Silver.
225 Bush St., Suite 1453,
San Francisco, CA 94104.
CANADA: Fenn Company, Inc. **DALLAS:** Griffith Group

Business Administration
Joachim P. Rosler, PUBLISHER
Marie M. Beaumonte, GENERAL MANAGER
Alyson M. Lane, BUSINESS MANAGER
Constance Holmes, MANAGER, ADVERTISING ACCOUNTING
AND COORDINATION

Chairman and Chief Executive Officer
John J. Hanley

Corporate Officers
Robert L. Biewen, Frances Newburg,
Joachim P. Rosler, VICE PRESIDENTS
Anthony C. Degutis, CHIEF FINANCIAL OFFICER

Program Development
Linnéa C. Elliott, DIRECTOR

Electronic Publishing
Martin O. K. Paul, DIRECTOR

Ancillary Products
Diane McGarvey, DIRECTOR

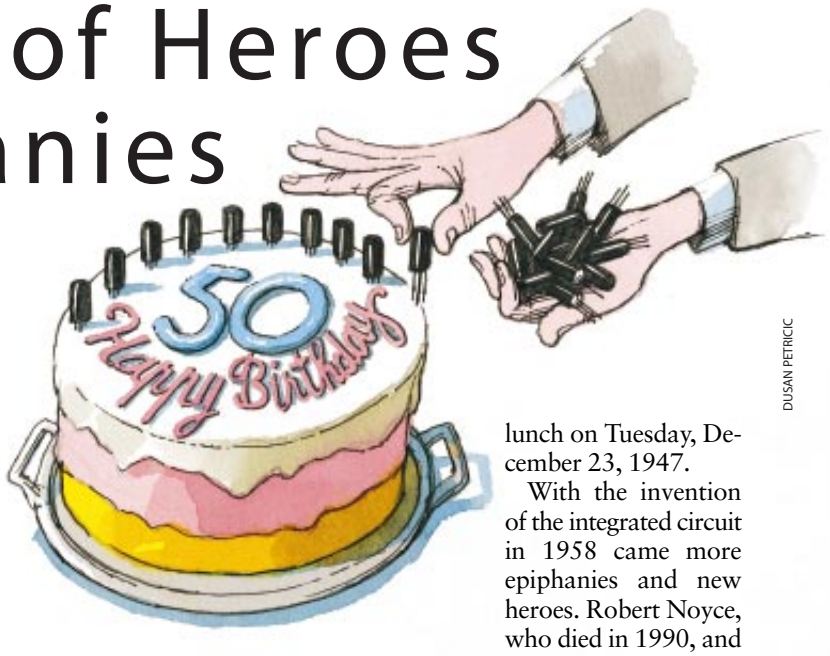
Scientific American, Inc.
415 Madison Avenue · New York, NY 10017-1111
(212) 754-0550

PRINTED IN U.S.A.

Copyright 1998 Scientific American, Inc.

Fifty Years of Heroes and Epiphanies

by Glenn Zorpette



DUSAN PETRICIC

Human beings crave legends, heroes and epiphanies. All three run through the history of solid-state electronics like special effects in one of Hollywood's summer blockbusters.

To begin with, solid state has an exceptionally poignant creation myth. Just after World War II, John Bardeen, a shy, quiet genius from a Wisconsin college town, and Walter Brattain, an ebullient, talkative experimenter raised in the backwoods of Washington State, assembled the most mundane of materials—a tiny slab of germanium, some bits of gold foil, a paper clip and some pieces of plastic—into a scraggly-looking gizmo. Ungainly as it was, the device was arguably one of the most beautiful things ever made. Every day of your life, you use thousands, if not millions, of its descendants.

After Bardeen and Brattain's achievement, their boss, the patrician William Shockley, improved on the delicate original device, making it more rugged and suitable for mass manufacture. What the three of them invented 50 years ago at Bell Telephone Laboratories was the transistor, the device that can switch an electric current on and off or take a minute current and amplify it into a much greater one. From its humble beginnings, the transistor has become the central, defining entity of the solid-state age, the ubiquitous sine qua non of just about every computer, data-handling appliance and power-amplifying circuit built since the 1960s.

"The Solid-State Century," as we have chosen to define it for this issue, extends from the work of Bardeen and company 50 years ago through whatever wonders the next 50 will surely bring. So far the first five decades have delivered not only the transistor but also the integrated circuit, in which millions of transistors are fabricated on tiny slivers of silicon; power transistors that can switch enormous flows of electric current; and optoelectronics, a huge category in its own right that includes the semiconductor lasers and detectors used in telecommunications and compact-disc systems.

In an attempt to impose order on such a mélange of marvels, we have divided this issue into three sections. The first covers devices—the transistor, semiconductor lasers and so on. Section two focuses on the integrated circuit. Section three describes some intriguing possibilities for the near future of electronics, especially in microprocessors and computers.

In the first section we start with the chilly, overcast afternoon when Bardeen and Brattain demonstrated their germanium-and-foil whatsit to suitably impressed executives at Bell Labs. Let's take a little license and say that the solid-state age was born right there and then, in Murray Hill, N.J., just after lunch on Tuesday, December 23, 1947.

With the invention of the integrated circuit in 1958 came more epiphanies and new heroes. Robert Noyce, who died in 1990, and Jack Kilby, who is profiled in this issue, separately conceived of integrating multiple transistors into a single, tiny piece of semiconductor material. As he recalls for interviewer Alan Goldstein, Kilby nurtured his idea in a laboratory that he had to himself for a hot summer month while his colleagues were all on vacation.

By the mid-1960s another hero, Gordon Moore (also profiled in this issue) noticed that the number of transistors that could be put on a chip was doubling every 12 months. (The doubling period has since lengthened to nearly two years.) Recently, however, some industry sages—including Moore himself—have begun openly speculating about when "Moore's Law" may finally come to an end and about what the industry will be like after it does. In this issue, we take up the subject in several articles, including "Technology and Economics in the Semiconductor Industry" and "Toward 'Point One.'"

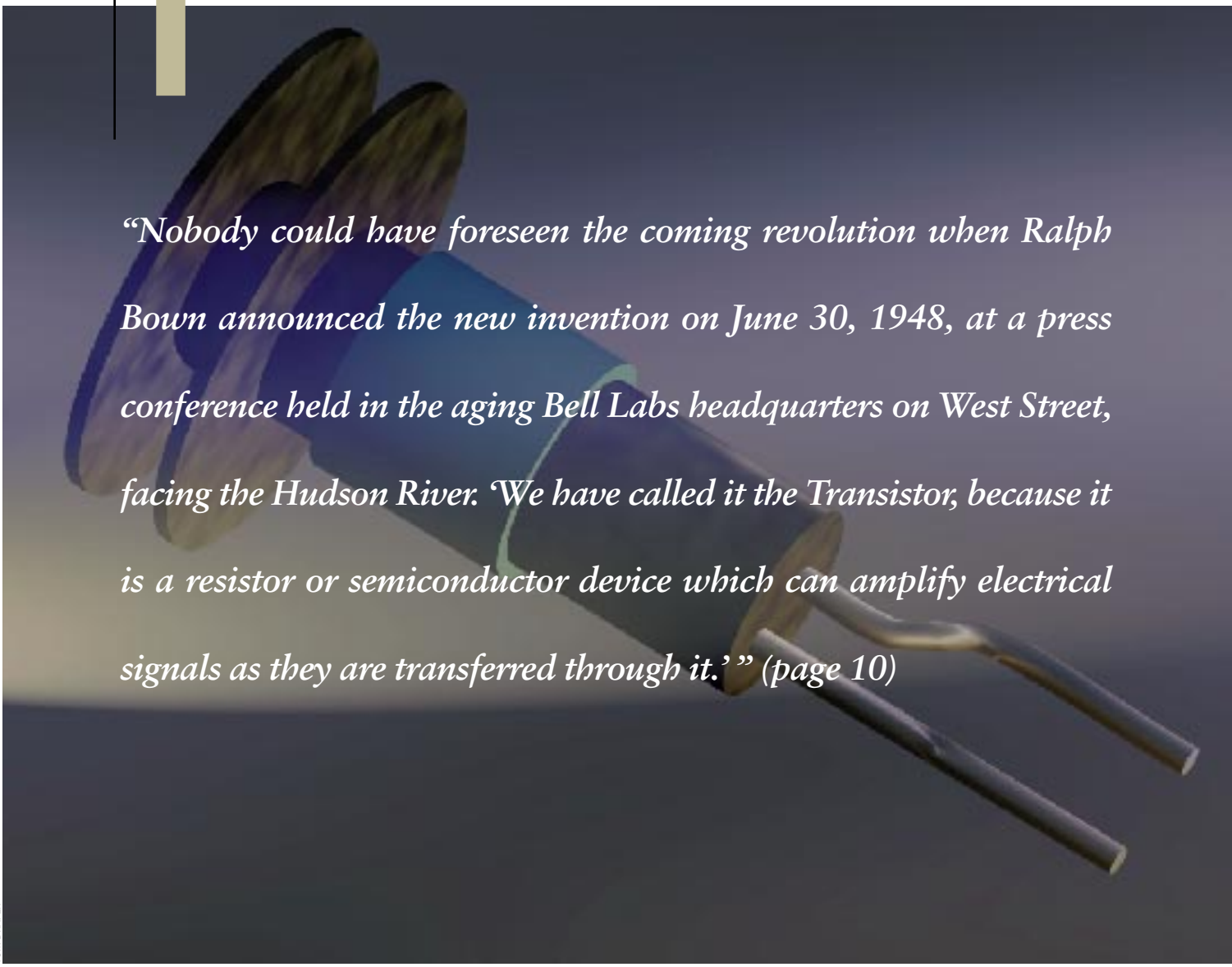
What it all comes down to, of course, are products. And extrapolating from past trends in the solid-state arena, the performance of some of them will truly astound. In "Microprocessors in 2020," David A. Patterson writes that it is not unreasonable to expect that two decades from now, a single desktop computer will be as powerful as all the computers in Silicon Valley today.

At the 50-year mark, the solid-state age has yet to show any sign of languor or dissipation in any of its categories. In microelectronics, chips with 10 million transistors are about to become available. In power electronics, a new type of device, the insulated gate bipolar transistor (IGBT) is revolutionizing the entire field. In optoelectronics, astonishing devices that exploit quantum effects are beginning to dominate. And it may not be too soon to identify a few new candidates for hero status—people such as the quantum-well wizard Federico Capasso of Lucent Technologies (which includes Bell Labs) and B. Jayant Baliga, the inventor of the IGBT, who describes his transistor in this issue. As we pass the halfway point in the solid-state century, it is clear that the cavalcade of legends, heroes and epiphanies is nowhere near over yet. SA

GLENN ZORPETTE is project editor for this special issue.

Copyright 1997 Scientific American, Inc.

1 The Transistor



"Nobody could have foreseen the coming revolution when Ralph Bown announced the new invention on June 30, 1948, at a press conference held in the aging Bell Labs headquarters on West Street, facing the Hudson River. 'We have called it the Transistor, because it is a resistor or semiconductor device which can amplify electrical signals as they are transferred through it.'" (page 10)

TOM DRAPER

Copyright 1998 Scientific American, Inc.

In December 1947 three researchers demonstrated a device that would change the way humankind works and plays

BIRTH OF AN ERA

by Michael Riordan and Lillian Hoddeson

William Shockley was extremely agitated. Speeding through the frosty hills west of Newark, N.J., on the morning of December 23, 1947, he hardly noticed the few vehicles on the narrow country road leading to Bell Telephone Laboratories. His mind was on other matters.

Arriving just after 7 A.M., Shockley parked his MG convertible in the company lot, bounded up two flights of stairs and rushed through the deserted corridors to his office. That afternoon his research team was to demonstrate a promising new electronic device to his boss. He had to be ready. An amplifier based on a semiconductor, he knew, could ignite a revolution. Lean and hawk-nosed, his temples graying and his thinning hair slicked back from a proud, jutting forehead, Shockley had dreamed of inventing such a device for almost a decade. Now his dream was about to come true.

About an hour later John Bardeen and Walter Brattain pulled up at this modern research campus in Murray Hill, 20 miles from New York City. Members of Shockley's solid-state physics group, they had made the crucial breakthrough a week before. Using little more than a tiny, nondescript slab of the element germanium, a thin plastic wedge and a shiny strip of gold foil, they had boosted an electrical signal almost 100-fold.

Soft-spoken and cerebral, Bardeen had come up with the key ideas, which were quickly and skillfully implemented by the genial Brattain, a salty, silver-haired man who liked to tinker with equipment almost as much as he loved to gab. Working shoulder to shoulder for most of the prior month, day after day except on Sundays, they had finally coaxed their curious-looking gadget into operation.

That Tuesday morning, while Bardeen completed a few calculations in his office, Brattain was over in his laboratory with a technician, making last-minute checks on their amplifier. Around one edge of a triangular plastic wedge, he had glued a small strip of gold foil, which he carefully slit along this edge with

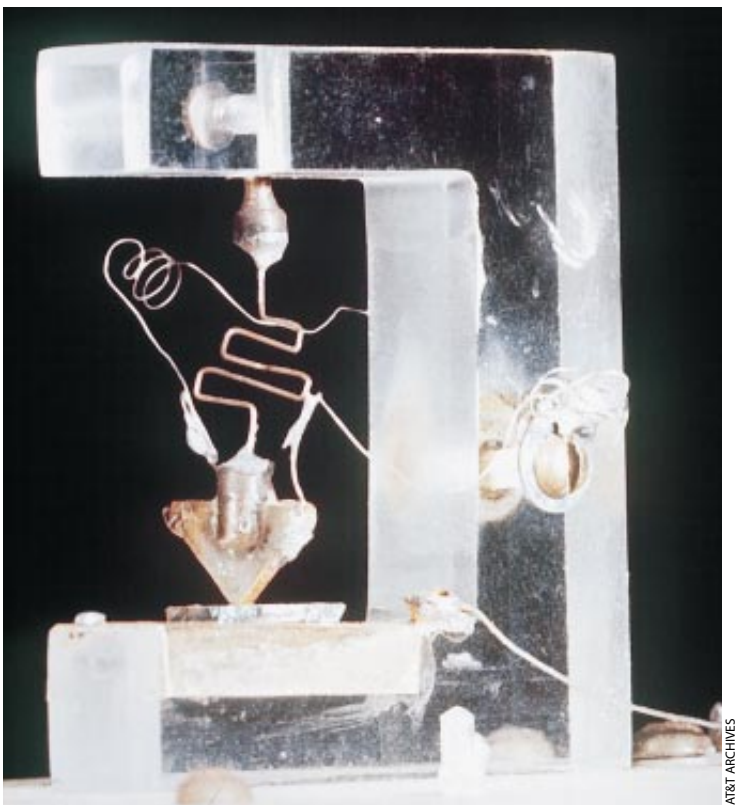
AT&T ARCHIVES

INVENTORS Shockley (*seated*), Bardeen (*left*) and Brattain (*right*) were the first to demonstrate a solid-state amplifier (*opposite page*).

Copyright 1997 Scientific American, Inc.

a razor blade. He then pressed both wedge and foil down into the dull-gray germanium surface with a makeshift spring fashioned from a paper clip. Less than an inch high, this delicate contraption was clamped clumsily together by a U-shaped piece of plastic resting upright on one of its two arms. Two copper wires soldered to edges of the foil snaked off to batteries, transformers, an oscilloscope and other devices needed to power the gadget and assess its performance.

Occasionally, Brattain paused to light a cigarette and gaze through blinds on the window of his clean, well-equipped lab. Stroking his mustache, he looked out across a baseball diamond on the spacious rural campus to a wooded ridge of the Watchung Mountains—worlds apart from the cramped, dusty laboratory he had occupied in downtown New York City before the war. Looking up, he saw slate-colored clouds



AT&T ARCHIVES

stretching off to the horizon. A light rain soon began to fall.

At age 45, Brattain had come a long way from his years as a roughneck kid growing up in the Columbia River basin. As a sharpshooting teenager, he helped his father grow corn and raise cattle on the family homestead in Tonasket, Wash., close to the Canadian border. "Following three horses and a harrow in the dust," he often joked, "was what made a physicist out of me."

Brattain's interest in the subject was sparked by two professors at Whitman College, a small liberal arts institution in the southeastern corner of the state. It carried him through graduate school at Oregon and Minnesota to a job in 1929 at Bell Labs, where he had remained—happy to be working at the best industrial research laboratory in the world.

Bardeen, a 39-year-old theoretical physicist, could hardly have been more different. Often lost in thought, he came across as very shy and self-absorbed. He was extremely parsimonious with his words, parceling them out softly in a deliberate monotone as if each were a precious gem never to be squandered. "Whispering John," some of his friends called him. But whenever he spoke, they listened. To many, he was an oracle.

Raised in a large academic family, the second son of the dean of the University of Wisconsin medical school, Bardeen had been intellectually precocious. He grew up among the ivied dorms and the sprawling frat houses lining the shores of Lake Mendota near downtown Madison, the state capital. Entering the university at 15, he earned two degrees in electrical engineering and worked a few years in industry before heading to Princeton University in 1933 to pursue a Ph.D. in physics.

In the fall of 1945 Bardeen took a job at Bell Labs, then winding down its wartime research program and gearing up for an expected postwar boom in electronics. He initially shared an office with Brattain, who had been working on semiconductors since the early 1930s, and Bardeen soon became intrigued by these curious materials, whose electrical properties were just beginning to be understood. Poles apart temperamentally, the two men became fast friends, often playing weekend golf together at the local country club.

Shortly after lunch that damp December day, Bardeen joined Brattain in his laboratory. Outside, the rain had changed over to snow, which was just beginning to accumulate. Shockley arrived about 10 minutes later, accompanied by his boss, acoustics expert Harvey Fletcher, and by Bell's research director, Ralph Bown—a tall, broad-shouldered man fond of expensive suits and fancy bow ties.

"The Brass," thought Bardeen a little contemptuously, using a term he had picked up from wartime work with the navy. Certainly these two executives would appreciate the commercial promise of this device. But could they really understand what was going on inside that shiny slab of germanium? Shockley might be comfortable rubbing elbows and bantering with the higher-ups, but Bardeen would rather be working on the physics he loved.

After a few words of explanation, Brattain powered up his equipment. The others watched the luminous spot that was racing across the oscilloscope screen jump and fall abruptly as he switched the odd contraption in and out of the circuit using a toggle switch. From the height of the jump, they could easily tell it was boosting the input signal many times whenever it was included in the loop. And yet there wasn't a single vacuum tube in the entire circuit!

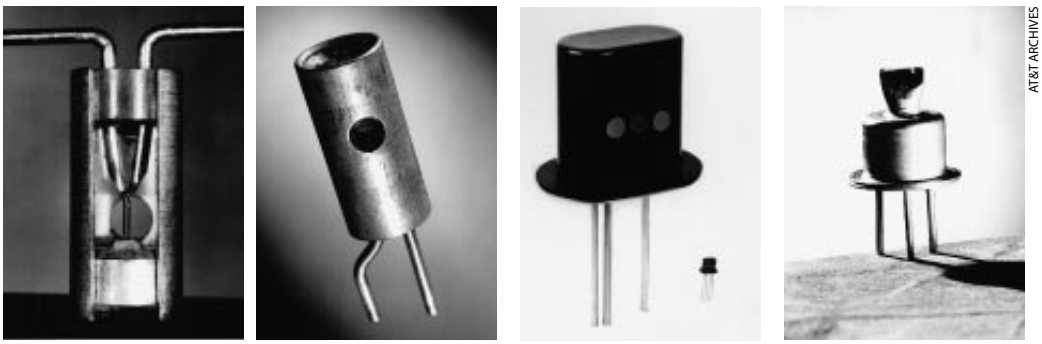
Then, borrowing a page from the Bell history books, Brattain spoke a few impromptu words into a microphone. They watched the sudden look of surprise on Bown's bespectacled face as he reacted to the sound of Brattain's gravelly voice booming in his ears through the headphones. Bown passed them to Fletcher, who shook his head in wonder shortly after putting them on.

For Bell Telephone Laboratories, it was an archetypal moment. More than 70 years earlier, a similar event had occurred in the attic of a boardinghouse in Boston, Mass., when Alexander Graham Bell uttered the words, "Mr. Watson, come here. I want you."

This article is excerpted from Crystal Fire: The Birth of the Information Age, *by Michael Riordan and Lillian Hoddeson. Copyright © 1997 by Michael Riordan and Lillian Hoddeson. Reprinted with permission of the publisher, W. W. Norton & Company, Inc.*

Copyright 1997 Scientific American, Inc.

Transistor Hall of Fame



AT&T ARCHIVES

Early transistors from Bell Laboratories were housed in a variety of ways. Shown here are point-contact transistors (*first two photographs from left*). The point-contact dates to 1948 and was essentially a packaged version of the original device demonstrated in 1947. Models from the late 1950s included the grown junction transistor (*second photograph from right*) and the diffused base transistor (*far right*).

In the weeks that followed, however, Shockley was torn by conflicting emotions. The invention of the transistor, as Bardeen and Brattain's solid-state amplifier soon came to be called, had been a "magnificent Christmas present" for his group and especially for Bell Labs, which had staunchly supported their basic research program. But he was chagrined to have had no direct role in this crucial breakthrough. "My elation with the group's success was tempered by not being one of the inventors," he recalled many years later. "I experienced frustration that my personal efforts, started more than eight years before, had not resulted in a significant inventive contribution of my own."

Wonderland World

Growing up in Palo Alto and Hollywood, the only son of a well-to-do mining engineer and his Stanford University–educated wife, Bill Shockley had been raised to consider himself special—a leader of men, not a follower. His interest in science was stimulated during his boyhood by a Stanford professor who lived in the neighborhood. It flowered at the California Institute of Technology, where he majored in physics before heading east in 1932 to seek a Ph.D. at the Massachusetts Institute of Technology. There he dived headlong into the Wonderland world of quantum mechanics, where particles behave like waves and waves like particles, and began to explore how streams of electrons trickle through crystalline materials such as ordinary table salt. Four years later, when Bell Labs lifted its Depression-era freeze on new employees, the cocky young Californian was the first new physicist to be hired.

With the encouragement of Mervin Kelly, then Bell's research director, Shockley began seeking ways to fashion a rugged solid-state device to replace the balky, unreliable switches and amplifiers commonly used in phone equipment. His familiarity with the weird quantum world gave him a decided advantage in this quest. In late 1939 he thought he had come up with a good idea—to stick a tiny bit of weathered copper screen inside a piece of semiconductor. Although skeptical, Brattain helped him build this crude device early the next year. It proved a complete failure.

Far better insight into the subtleties of solids was needed—and much purer semiconductor materials, too. World War II interrupted Shockley's efforts, but wartime research set the stage for major breakthroughs in electronics and communications once the war ended. Stepping in as Bell Labs vice president, Kelly recognized these unique opportunities and organized a solid-state physics group, installing his ambitious protégé as its co-leader.

Shockley's elation was tempered by not being one of the inventors.

Soon after returning to the labs in early 1945, Shockley came up with another design for a semiconductor amplifier. Again, it didn't work. And he couldn't understand why. Discouraged, he turned to other projects, leaving the conundrum to Bardeen and Brattain. In the course of their research, which took almost two years, they stumbled on a different—and successful—way to make such an amplifier.

Their invention quickly spurred Shockley into a bout of feverish activity. Galled at being upstaged, he could think of little else besides semiconductors for over a month. Almost every moment of free time he spent on trying to design an even better solid-state amplifier, one that would be easier to manufacture and use. Instead of whooping it up with other scientists and engineers while attending two conferences in Chicago, he spent New Year's Eve cooped up in his hotel room with a pad and a few pencils, working into the early-morning hours on yet another of his ideas.

By late January 1948 Shockley had figured out the important details of his own design, filling page after page of his lab notebook. His approach would use nothing but a small strip of semiconductor material—silicon or germanium—with three wires attached, one at each end and one in the middle. He eliminated the delicate "point contacts" of Bardeen and Brattain's unwieldy contraption (the edges of the slit gold foil wrapped around the plastic wedge). Those, he figured, would make manufacturing difficult and lead to quirky performance. Based on boundaries or "junctions" to be established within the semiconductor material itself, his amplifier should be much easier to mass-produce and far more reliable.

But it took more than two years before other Bell scientists perfected the techniques needed to grow germanium crystals with the right characteristics to act as transistors and amplify electrical signals. And not for a few more years could such "junction transistors" be produced in quantity. Meanwhile Bell engineers plodded ahead, developing point-contact transistors based on Bardeen and

Copyright 1997 Scientific American, Inc.

Brattain's ungainly invention. By the middle of the 1950s, millions of dollars in new equipment based on this device was about to enter the telephone system.

Still, Shockley had faith that his junction approach would eventually win out. He had a brute confidence in the superiority of his ideas. And rarely did he miss an opportunity to tell Bardeen and Brattain, whose relationship with their abrasive boss rapidly soured. In a silent rage, Bardeen left Bell Labs in 1951 for an academic post at the University of Illinois. Brattain quietly got himself reassigned elsewhere within the labs, where he could pursue research on his own. The three men crossed paths again in Stockholm, where they shared the 1956 Nobel Prize for Physics for their invention of the transistor. The tension eased a bit after that—but not much.

By the mid-1950s physicists and electrical engineers may have recognized the transistor's significance, but the general public was still almost completely oblivious. The millions of radios, television sets and other electronic devices produced every year by such gray-flannel giants of American industry as General Electric, Philco, RCA and Zenith came in large, clunky boxes powered by balky vacuum tubes that took a minute or so to warm up before anything could happen. In 1954 the transistor was largely perceived as an expensive laboratory curiosity with only a few specialized applications, such as hearing aids and military communications.

But that year things started to change dramatically. A small, innovative Dallas company began producing junction transistors for portable radios, which hit U.S. stores at $49.95. Texas Instruments curiously abandoned this market, only to see it cornered by a tiny, little known Japanese company called Sony. Transistor radios you could carry around in your shirt pocket soon became a minor status symbol for teenagers in the suburbs sprawling across the American landscape. After Sony started manufacturing TV sets powered by transistors in the 1960s, U.S. leadership in consumer electronics began to wane.

Vast fortunes would eventually be made in an obscure valley south of San Francisco, then filled with apricot orchards. In 1955 Shockley left Bell Labs for northern California, intent on making the millions he thought he deserved, founding the first semiconductor company in the valley. He lured top-notch scientists and engineers away from Bell and other companies, ambitious men like himself who soon jumped ship to start their own firms. What became famous around the world as Silicon Valley began with Shockley Semiconductor Laboratory, the progenitor of hundreds of companies like it, a great many of them far more successful.

The transistor has indeed proved to be what Shockley so presciently called the "nerve cell" of the Information Age. Hardly a unit of electronic equipment can be made today without it. Many thousands—and even millions—of them are routinely packed with other microscopic specks onto slim crystalline slivers of silicon called microprocessors, better known as microchips. By 1961 transistors were the foundation of a $1-billion semiconductor industry whose sales were doubling almost every year. Over three decades later, the computing power that had once required rooms full of bulky, temperamental electronic equipment is now easily loaded into

JASON GOLTZ

RADIOS went from living rooms to jacket pockets in the early 1960s, not long after the appearance of the first transistor-based units. Small radios soon became a status symbol among teenagers and young adults. Integrated circuits have permitted even smaller personal systems.

ARCHIVE PHOTOS/HIRZ

ARCHIVE/HERBERT

Birth of an Era

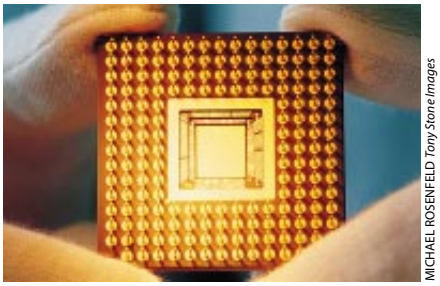
Copyright 1997 Scientific American, Inc.

units that can sit on a desktop, be carried in a briefcase or even rest in the palm of one's hand. Words, numbers and images flash around the globe almost instantaneously via transistor-equipped satellites, fiber-optic networks, cellular telephones and facsimile machines.

Through their landmark efforts, Bardeen, Brattain and Shockley had struck the first glowing sparks of a great technological fire that has raged through the rest of the century and shows little sign of abating. Cheap, portable and reliable equipment based on transistors can now be found in almost every village and hamlet in the world. This tiny invention has made the world a far smaller and more intimate place than ever before.

ARCHIVE PHOTOS

ARCHIVE PHOTOS


MICHAEL ROSENFELD *Tony Stone Images*

ARCHIVE PHOTOS

DISSEMINATION OF INFORMATION has been transformed by the integration of transistors onto chips (*above, top*). Computers that are inexpensive, small and rugged (*right*) in comparison with their predecessors (*above*) are now able to tap into global-spanning networks. They supplement more traditional conveyors of information (*left*), including the one the reader is now holding.

The Media Yawns

Nobody could have foreseen the coming revolution when Ralph Bown announced the new invention on June 30, 1948, at a press conference held in the aging Bell Labs headquarters on West Street, facing the Hudson River opposite the bustling Hoboken Ferry. "We have called it the Transistor," he began, slowly spelling out the name, "because it is a resistor or semiconductor device which can amplify electrical signals as they are transferred through it." Comparing it to the bulky vacuum tubes that served this purpose in virtually every electrical circuit of the day, he told reporters that the transistor could accomplish the very same feats and do them much better, wasting far less power.

But the press paid little attention to the small cylinder with two flimsy wires poking out of it that was being demonstrated by Bown and his staff that sweltering summer day. None of the reporters suspected that the physical process silently going on inside this innocuous-looking metal tube, hardly bigger than the rubber erasers on the ends of their pencils, would utterly transform their world.

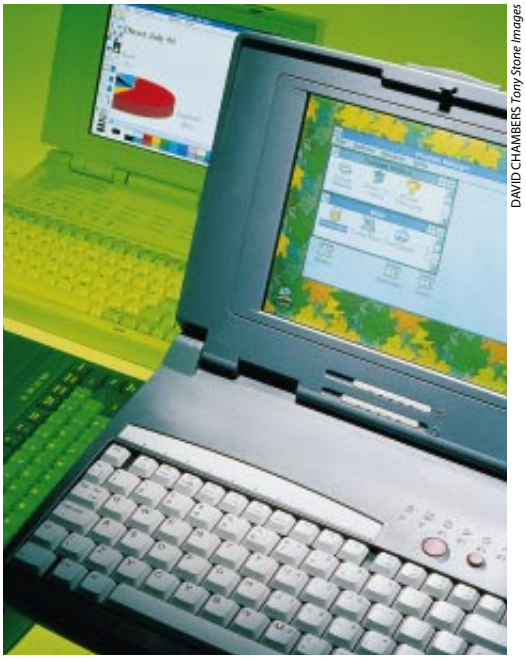
Editors at the *New York Times* were intrigued enough to mention the breakthrough in the July 1 issue, but they buried the story on page 46 in "The News of Radio." After noting that *Our Miss Brooks* would replace the regular CBS Monday-evening program *Radio Theatre* that summer, they devoted a few paragraphs to the new amplifier.

"A device called a transistor, which has several applications in radio where a vacuum tube ordinarily is employed, was demonstrated for the first time yesterday at Bell Telephone Laboratories," began the piece, noting that it had been employed in a radio receiver, a telephone system and a television set. "In the shape of a small metal cylinder about a half-inch long, the transistor contains no vacuum, grid, plate or glass envelope to keep the air away," the column continued. "Its action is instantaneous, there being no warm-up delay since no heat is developed as in a vacuum tube."

Perhaps too much other news was breaking that sultry Thursday morning. Turnstiles on the New York subway system, which until midnight had always droned to the dull clatter of nickels, now marched only to the music of dimes. Subway commuters responded with resignation. Idlewild Airport had opened for business the previous day in the swampy meadowlands just east of Brooklyn, supplanting La Guardia as New York's principal destination for international flights. And the hated Boston Red Sox had beaten the world champion Yankees 7 to 3.

Earlier that week the gathering clouds of the cold war had darkened dramatically over Europe after Soviet occupation forces in eastern Germany refused to allow Allied convoys to carry any more supplies into West Berlin. The U.S. and Britain responded to this blockade with a massive airlift. Hundreds of transport planes brought the thousands of tons of food and fuel needed daily by

Copyright 1997 Scientific American, Inc.



DAVID CHAMBERS *Tony Stone Images*

the more than two million trapped citizens. All eyes were on Berlin. "The incessant roar of the planes—that typical and terrible 20th Century sound, a voice of cold, mechanized anger—filled every ear in the city," *Time* reported. An empire that soon encompassed nearly half the world's population seemed awfully menacing that week to a continent weary of war.

To almost everyone who knew about it, including its two inventors, the transistor was just a compact, efficient, rugged replacement for vacuum tubes. Neither Bardeen nor Brattain foresaw what a crucial role it was about to play in computers, although Shockley had an inkling. In the postwar years electronic digital computers, which could then be counted on the fingers of a single hand, occupied large rooms and required teams of watchful attendants to replace the burned-out elements among their thousands of overheated vacuum tubes. Only the armed forces, the federal government and major corporations could afford to build and operate such gargantuan, power-hungry devices.

Five decades later the same computing power is easily crammed inside a pocket calculator costing around $10, thanks largely to microchips and the transistors on which they are based. For the amplifying action discovered at Bell Labs in 1947–1948 actually takes place in just a microscopic sliver of semiconductor material and—in stark contrast to vacuum tubes—produces almost no wasted heat. Thus, the transistor has lent itself readily to the relentless miniaturization and the fantastic cost reductions that have put digital computers at almost everybody's fingertips. Without the transistor, the personal computer would have been inconceivable, and the Information Age it spawned could never have happened.

Linked to a global communications network that has itself undergone a radical transformation because of transistors, computers are now revolutionizing the ways we obtain and share information. Whereas our parents learned about the world by reading newspapers and magazines or by listening to the baritone voice of Edward R. Murrow on their radios, we can now access far more information at the click of a mouse—and from a far greater variety of sources. Or we witness such earth-shaking events as the fall of the Soviet Union in the comfort of our living rooms, often the moment they occur and without interpretation.

Although Russia is no longer the looming menace it was during the cold war, nations that have embraced the new information technologies based on transistors and microchips have flourished. Japan and its retinue of developing eastern Asian countries increasingly set the world's communications standards, manufacturing much of the necessary equipment. Television signals penetrate an ever growing fraction of the globe via satellite. Banks exchange money via rivers of ones and zeroes flashing through electronic networks all around the world. And boy meets girl over the Internet.

No doubt the birth of a revolutionary artifact often goes unnoticed amid the clamor of daily events. In half a century's time, the transistor, whose modest role is to amplify electrical signals, has redefined the meaning of power, which today is based as much on the control and exchange of information as it is on iron or oil. The throbbing heart of this sweeping global transformation is the tiny solid-state amplifier invented by Bardeen, Brattain and Shockley. The crystal fire they ignited during those anxious postwar years has radically reshaped the world and the way its inhabitants now go about their daily lives. SA

The Authors

MICHAEL RIORDAN and LILLIAN HODDESON are co-authors of *Crystal Fire: The Birth of the Information Age*. Riordan is the assistant to the director of the Stanford Linear Accelerator Center and a research physicist at the University of California, Santa Cruz. He holds two degrees in physics from the Massachusetts Institute of Technology and is co-author of *The Solar Home Book* and *The Hunting of the Quark*. Hoddeson, an associate professor of history at the University of Illinois at Urbana-Champaign, is co-author of *The Birth of Particle Physics* and co-author, with Vicki Daitch, of the forthcoming *Gentle Genius: The Life and Science of John Bardeen.*

Further Reading

THE PATH TO THE CONCEPTION OF THE JUNCTION TRANSISTOR. William Shockley in *IEEE Transactions on Electron Devices*, Vol. ED-23, No. 7, pages 597–620; July 1976.

REVOLUTION IN MINIATURE: THE HISTORY AND IMPACT OF SEMICONDUCTOR ELECTRONICS. Ernest Braun and Stuart MacDonald. Cambridge University Press, 1978.

AN AGE OF INNOVATION: THE WORLD OF ELECTRONICS 1930–2000. The editors of *Electronics* magazine. McGraw-Hill, 1981.

A HISTORY OF ENGINEERING AND SCIENCE IN THE BELL SYSTEM, Vol. 4: PHYSICAL SCIENCES and Vol. 6: ELECTRONICS TECHNOLOGY. Edited by technical staff, AT&T Bell Laboratories. AT&T Bell Laboratories, 1983.

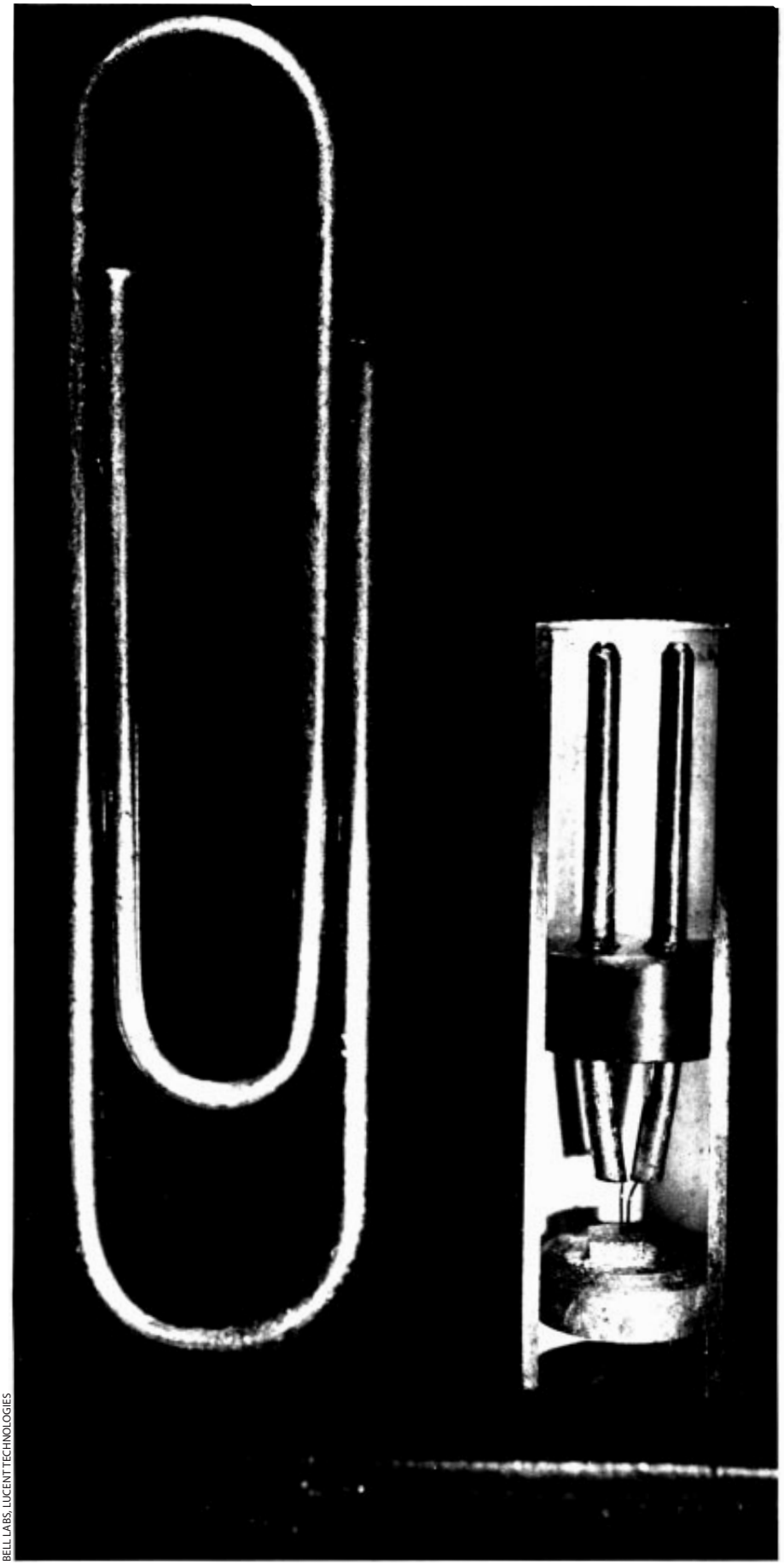
THE ORIGIN, DEVELOPMENT, AND PERSONALITY OF MICROELECTRONICS. R. M. Warner in *Transistors: Fundamentals for the Integrated-Circuit Engineer*. John Wiley & Sons, 1983.

ENGINEERS AND ELECTRONS. John D. Ryder and Donald G. Fink. IEEE Press, 1984.

AMERICAN GENESIS: A CENTURY OF INVENTION AND TECHNOLOGICAL ENTHUSIASM. Thomas P. Hughes. Penguin Books, 1990.

CRYSTALS, ELECTRONS AND TRANSISTORS. Michael Eckert and Helmut Shubert. AIP Press, 1990.

Copyright 1997 Scientific American, Inc.



BELL LABS, LUCENT TECHNOLOGIES

PAPER CLIP AND TRANSISTOR are compared to emphasize the transistor's size. Metal tube of transistor is cut away. Germanium crystal is tiny block on disk at bottom. Two cat's whiskers are mounted on heavy leads.

This article, which appeared in the September 1948 issue of SCIENTIFIC AMERICAN, *offered one of the earliest surveys of transistor technology. It is reprinted here in its original form.*

IN 1906 a young American electrical engineer named Lee De Forest discovered that if an electrified wire grid was placed across the path of a stream of electrons in a vacuum tube, the flow of electrons could be controlled in some rather interesting ways. The flow could be interrupted, reduced or stopped entirely; a feeble current of electrons entering at one end of the tube could be "amplified" to a powerful current at the outgoing end. It was this classically simple invention by De Forest that gave birth to the tremendous technology of electronics. From it came radio, television, radar, X-ray cameras, electron microscopes, guided missiles, electronic calculators, robot machine tenders, electronic burglar alarms, instruments that examine materials for invisible flaws, doors that open themselves —and doubtless greater wonders are yet to come. The electronic tube is easily one of the most ingenious inventions and most versatile tools of our fabulous century.

Since De Forest's elementary discovery, the electronic tube has been developed enormously. Electronic theory also has advanced rapidly, and it now appears that the vacuum tube is far from the last word. Within the past few months a group of physicists at the Bell Telephone Laboratories has made another profound and simple finding which may rank in importance with that of De Forest. In essence, it is a method of controlling electrons in a solid crystal instead of in a vacuum. This discovery has yielded a device called the transistor (so named because it transfers an electrical signal across a resistor) which can do many of the things that a vacuum tube does. Indeed, it has certain advantages over the vacuum tube. It reduces the complicated, delicate tube to a simple rig consisting basically of a couple of fine wires—cat's whiskers in the radioman's language—and a small crystal; no vacuum is needed. The transistor does not need to heat up, as a vacuum tube does, and so it goes to work instantly. It operates on a tiny amount of power—about one tenth of that used by an ordinary flashlight bulb. And it can be made almost vanishingly small. The present experimental model is about the size of the eraser on the end of a pencil.

The technological fruits of this inven-

Copyright 1997 Scientific American, Inc.

THE TRANSISTOR

Basic research in the electrical properties of solids has opened up an entirely new way of manipulating electrons to do useful work

by Frank H. Rockett

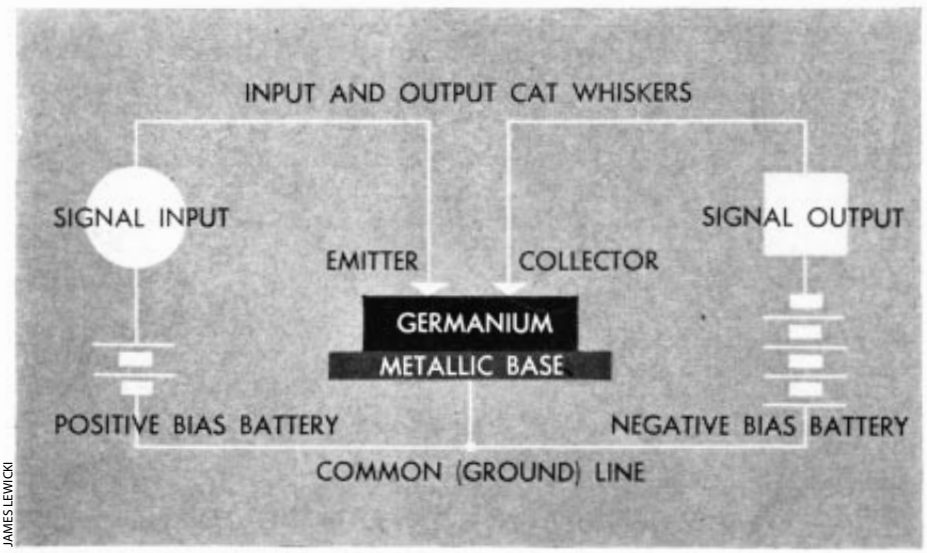
tion already appear extensive. The size of vacuum tubes is an important consideration in electronics, for it largely determines the size of the apparatus in which they are used. A television receiver requires about two dozen tubes; the celebrated computing machine at the University of Pennsylvania known as ENIAC has 18,000. With ingenuity and painstaking labor "subminiature" vacuum tubes only an inch long have been produced for some special purposes, but the transistor promises to reduce electronic equipment in general to an even smaller scale. Not only is the transistor itself tiny, but it needs so little power, and uses that little so efficiently (as a radio amplifier its efficiency is 25 per cent, against a vacuum tube's 10 per cent) that the size of batteries needed to operate portable devices can also be reduced. Thus the transistor makes possible tinier hearing aids, really small portable radios, more compact electronic devices for aircraft and a great reduction in the bulk of stationary equipment. In combination with printed circuits—the compact new wiring system—it may open up entirely new applications for electronics. The transistor also suggests the possibility of a considerable improvement in telephone transmission, because amplifiers for long-distance cables can be built small and mounted inconspicuously on telephone poles, and a miniature amplifier may even be built into the telephone receiver to strengthen weak signals.

Beyond this, the transistor vastly simplifies the manufacture and maintenance of electronic equipment. Because of its simple, sturdy construction, it will be longer-lived and possibly less costly than vacuum tubes. The transistor has important limitations: its power output in the present research stage is small (a maximum of about one fortieth of a watt), and the highest frequency at which it can operate is about 10 megacycles (10 million cycles per second). But its power and frequency range are sufficient for most purposes in the regular broadcast, television and short-wave regions of the radio spectrum.

The transistor is the unexpected product of purely scientific curiosity. To understand how it was conceived and how it works one must examine the functions of the electronic tube and the way in which electric current is conducted by a solid. The basic purposes of an electronic tube are to convert an alternating current into a direct current (called rectification), to amplify the signal, to break it up into pulses instead of a continuous wave, or to make it oscillate, *i.e.*, beat in a regular rhythm at a calculated frequency. The tube itself was invented in 1905 by the English physicist J. Ambrose Fleming, who observed that if an alternating current was passed to a filament inside a vacuum tube the electrons would boil off the end of the filament as free particles and would travel across the vacuum to a positively charged plate at the other end. (This phenomenon, known as the Edison effect, had been noticed much earlier by Thomas Edison, but he had been unable to explain it and had made no practical use of it.) As long as the current was kept on, the electrons would move only towards the attracting positive plate; hence the tube was an easy means of changing alternating current into a direct, one-way signal. Fleming's tube, called a diode because it had two electrodes—the filament and the plate—could be used as a detector for radio signals. But De Forest's addition of the grid to control the electrons, making the tube a triode, was the step which gave the tube its great versatility and usefulness. Now the signal could be controlled and amplified (since a small number of electrons on the grid governed the flow of a much larger number from the filament to the plate). It could also be modified in other ways.

When not in a vacuum, electrons obviously are much less easy to control, since they cling more or less firmly to orbits about the nuclei of atoms. Whether a solid will conduct electricity depends on the degree of freedom of its electrons. Copper, a good conductor, has a single electron in its outer orbit or shell, and this relatively free electron readily serves to carry current. Most metals have such loosely held electrons, hence are good conductors. On the other hand, an element such as sulphur, whose electrons are all locked in place by tight bonds with the nucleus and with other atoms, does not conduct electricity; it is an insulator.

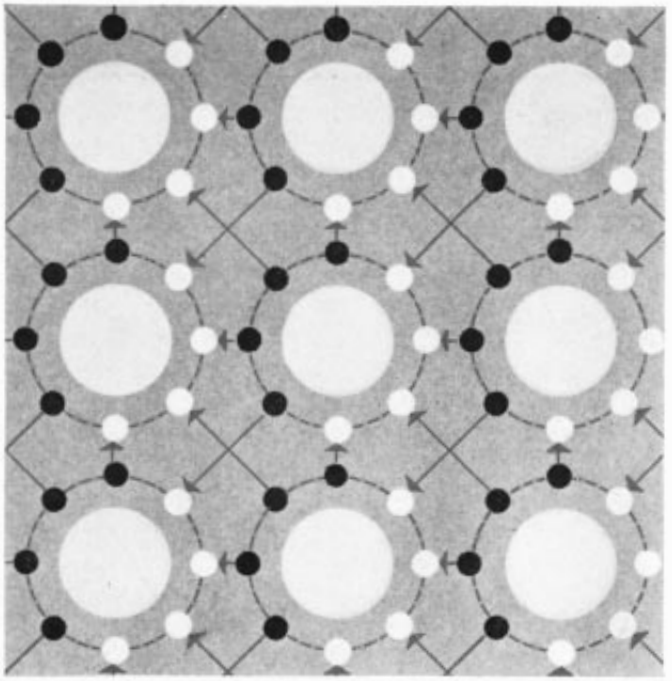
BETWEEN these extremes there is a class of materials known as semiconductors which furnish an occasional free electron for carrying current. Silicon and germanium are examples; they have about one free electron for every thousand atoms (as contrasted with copper, which has one



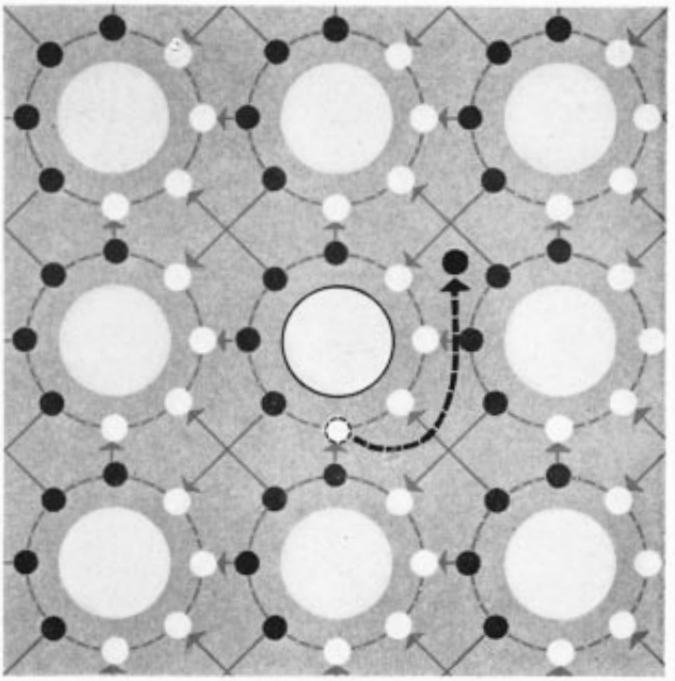
JAMES LEWICKI

SIMPLE CIRCUIT uses transistor in place of three-element vacuum tube. When small signal passes over surface of transistor between cat's whiskers, larger current passed through transistor is modulated in replica of small.

Copyright 1997 Scientific American, Inc.



ATOMS OF GERMANIUM, shown here in schematic crystal lattice, have four electrons *(black dots)* in outer orbits. These move freely into gaps in adjacent orbits.



ATOM OF PHOSPHORUS, introduced into germanium crystal, has one more outer electron. This is free to travel through the crystal, improving its conductivity.

for every atom). These semiconductors have long possessed a special interest for electronic researchers. The important fact about them is that the number of current-carrying electrons in them can be controlled. They can be made to act as conductors under some conditions and as insulators under others. Indeed, they are so sensitive that the current flowing in a semiconductor can be controlled by the brightness of a light shining on it in a region where a fine wire touches it. So this class of materials has been adapted to many uses. The crystal detector, used in early radios and now employed in an improved form in radar sets, is a semiconductor.

It was research into some of the mysterious electrical properties of semiconductors that led to the development of the transistor. It is helpful to try to visualize the electrical behavior of one of these substances. Picture a crystal of silicon (or germanium), which has four electrons in its outer shell—so-called valence electrons that hook the atoms together. Because they are fully occupied in forming bonds between the atoms, the electrons are not available for carrying electricity. Now suppose some impurity which has five valence electrons, say an atom of phosphorus, gets into the crystal. Four of these electrons become busy forming bonds with silicon atoms, but the fifth is free to carry current.

A more interesting case, and the one with which we are chiefly concerned here, is an impurity with three valence electrons, such as boron. One of the bonds needed for union with the silicon atoms is missing. The result is a state of disequilibrium, as the physicists say; there is some shifting around of bonds, but however they arrange themselves there is bound to be a missing electron. Because it is much easier to consider the movements of the gap created by the single missing electron than to follow the movements of the numerous other electrons as they create and fill in gaps, the missing electron is treated as an actual physical entity, though it is called a "hole." It has all the properties of an electron, such as mass and charge, except that, being the absence of an electron, its charge is positive instead of negative.

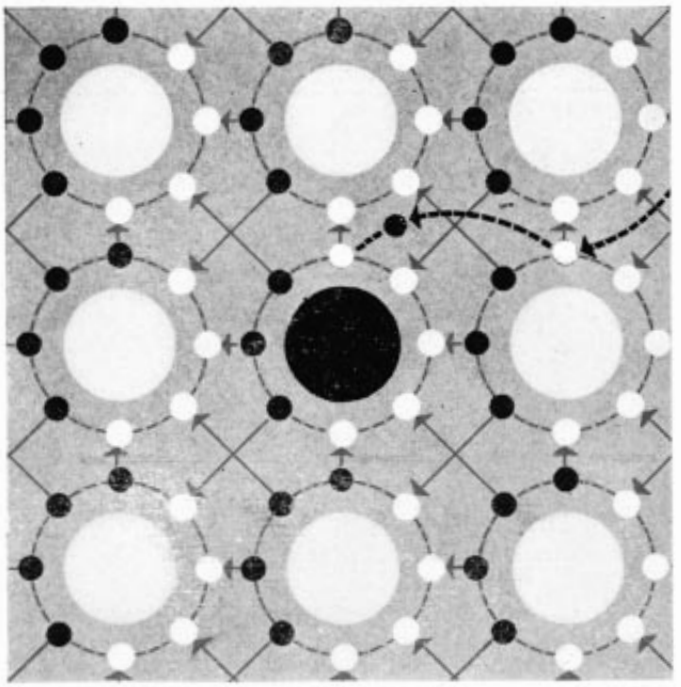
This, then, is a rough picture of the theory: the ability of a crystal semiconductor to conduct electricity is due to the presence of impurities that free some of the electrons which would otherwise be occupied in linking atoms. But a physicist at the Bell Telephone Laboratories, John Bardeen, became curious about a phenomenon that seemed to leave a hole in the theory. When a semiconductor is placed between two metallic contacts in an electrical circuit, one of the contacts being the point of a fine wire and the other a metal block, the arrangement acts as a rectifier, in a manner somewhat similar to the electronic tube. The reason is that the point contact between the semiconductor and the cat's whisker has a lower resistance to electrical flow in one direction than in the other. This difference in resistance accounts for the rectifying action of a crystal. Because it passes current predominantly in the direction of low resistance, the alternating voltage is converted to direct current.

One would suppose that the respective resistances to current flow in one direction or the other would vary with the physical properties or resistance of the materials forming the contacts. But experiments showed that the properties of the metals made much less difference than the theory had predicted. Bardeen decided that something must take place at the surface of the crystal that the theory had not explained. Aided by previous work on a similar problem by William Shockley, director of the semiconductor research at Bell Labs, Bardeen undertook a theoretical study of the conditions at the surface of a semiconductor.

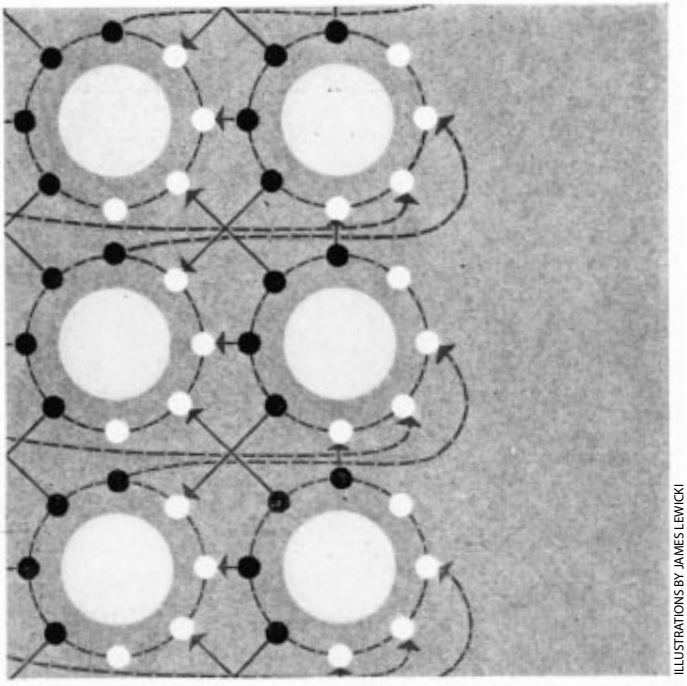
THE RESULT of this study was an important modification of the theory, which subsequent experiments were to prove correct. Bardeen reasoned that there were localized states on the surface of a semiconductor which differed from those in the interior. The number of such states, he said, was equal to the number of surface atoms. Like impurities in a crystal, these states produced holes capable of carrying current. These holes consisted of spaces on the exposed side of the atoms, which normally would be filled by electrons from adjacent atoms. This is an oversimplified picture of the theory, but it helps to make clear the essential concept: that the surface of a semiconductor is a better carrier of electricity than its interior. And Bardeen's theory satisfactorily accounted for the fact that the rectifying action of a crystal was independent of the particular metal used for the cat's whisker.

Shockley soon carried out an experiment that gave strong support to the theory. He reasoned that an externally

Copyright 1997 Scientific American, Inc.



ATOM OF BORON, also introduced into germanium, has one less outer electron. This "hole" is able to migrate through crystal in much the same way as a real electron.



SURFACE ATOMS, one side of which does not adjoin other atoms, have unfilled holes. These make surface of crystal a conductor of tiny currents that pass across it.

ILLUSTRATIONS BY JAMES LEWICKI

applied electric field should increase the conductivity of a crystal by inducing electrons out of the bonds. He placed a sheet of germanium in an intense electric field. The increase in conductivity of the germanium turned out to be less than the old theory predicted. But the measurements fitted in well with Bardeen's new theory. They could be explained by the assumption, suggested by his theory, that the conductive layer of electrons or holes on the surface of the germanium acted as a shield against penetration of the material by the electric field, just as metallic shields around parts of radio sets keep away stray electric fields.

Bardeen, Shockley and a colleague, W. H. Brattain, proceeded to further experiments and calculations, each new experiment resulting in refinements of the theory. They concluded that the superior conductivity of the surface layer of germanium in their experiments was accounted for chiefly by the presence of holes, and that these holes were produced not only by impurities and surface states but also by the current passing through the crystal.

These studies, indicating a method of controlling the electrons or holes in a crystal, led Bardeen and Brattain to the invention of the transistor. The device consists of two fine tungsten wires of which the tips, only two thousandths of an inch apart, rest on a germanium crystal soldered in turn to a metal disk. All these elements are housed in a metal cylinder which is connected electrically to the metal disk and crystal, thus forming the ground terminal. The cat's whisker wires are connected to pins that can be plugged into a socket (*see diagram on page 19*).

An electrical signal, modified by a small positive "bias" voltage to place it in the proper state for action on the crystal, is transmitted to the crystal by one of the cat's whiskers, called the emitter. The current releases holes in the surface layer of the crystal. The positive holes, flowing over the surface, are attracted to the second cat's whisker, which is biased negatively. The first whisker controls the number of holes flowing to the second whisker, in the same way that a vacuum tube grid controls the number of electrons flowing to the plate. The second whisker, called the collector, absorbs the current carried by the holes and passes on the signal, amplified 100 times. The amplification is partly due to the fact that a change in the incoming current to the crystal produces a greater change in the outgoing current. Most of it derives, however, from the great difference in resistance between the input and output ends of the circuit. The behavior of the electrons or holes is controlled in the crystal by superimposing variations on the positive and negative biased voltages applied to the emitter and collector. Thus the transistor is essentially a triode form of the well-known crystal diode detector used in radio.

Engineers at the Bell Laboratories have demonstrated that the transistor can be used as a voice amplifier, a television picture amplifier, a pulse amplifier and an oscillator. They have even produced a superheterodyne radio receiving set operating completely without vacuum tubes. Transistors were used in the set's amplifiers and in the local oscillator; conventional germanium crystal detectors served as mixer and detector, and selenium rectifiers were used in the power supply. This set performed as well as a conventional five-tube superheterodyne receiver. Since there are no vacuum tubes to heat up, a program comes in at full strength as soon as the set is switched on.

This instant response of the transistor is especially useful in pulse-type communication systems and in electronic computers. The transistor will also have a special value in electronic equipment that must operate continuously even when there are power failures. Such equipment, which includes telephone repeaters, fire and burglar alarms and the like, is generally equipped with batteries for an emergency power supply. The small power requirements of the transistor will make it possible to use batteries for a prolonged period.

YET all these applications are less important than the fundamental new knowledge that has been gained about the structure and energy states of solid matter and the electrical behavior of the surface atoms in a semiconductor. Basic study of these phenomena has been undertaken not only at Bell Labs but at Purdue University, the University of Pennsylvania and the Radiation Laboratory at the Massachusetts Institute of Technology. The holes in the crystal lattice of atoms obviously are a promising subject for further investigation.

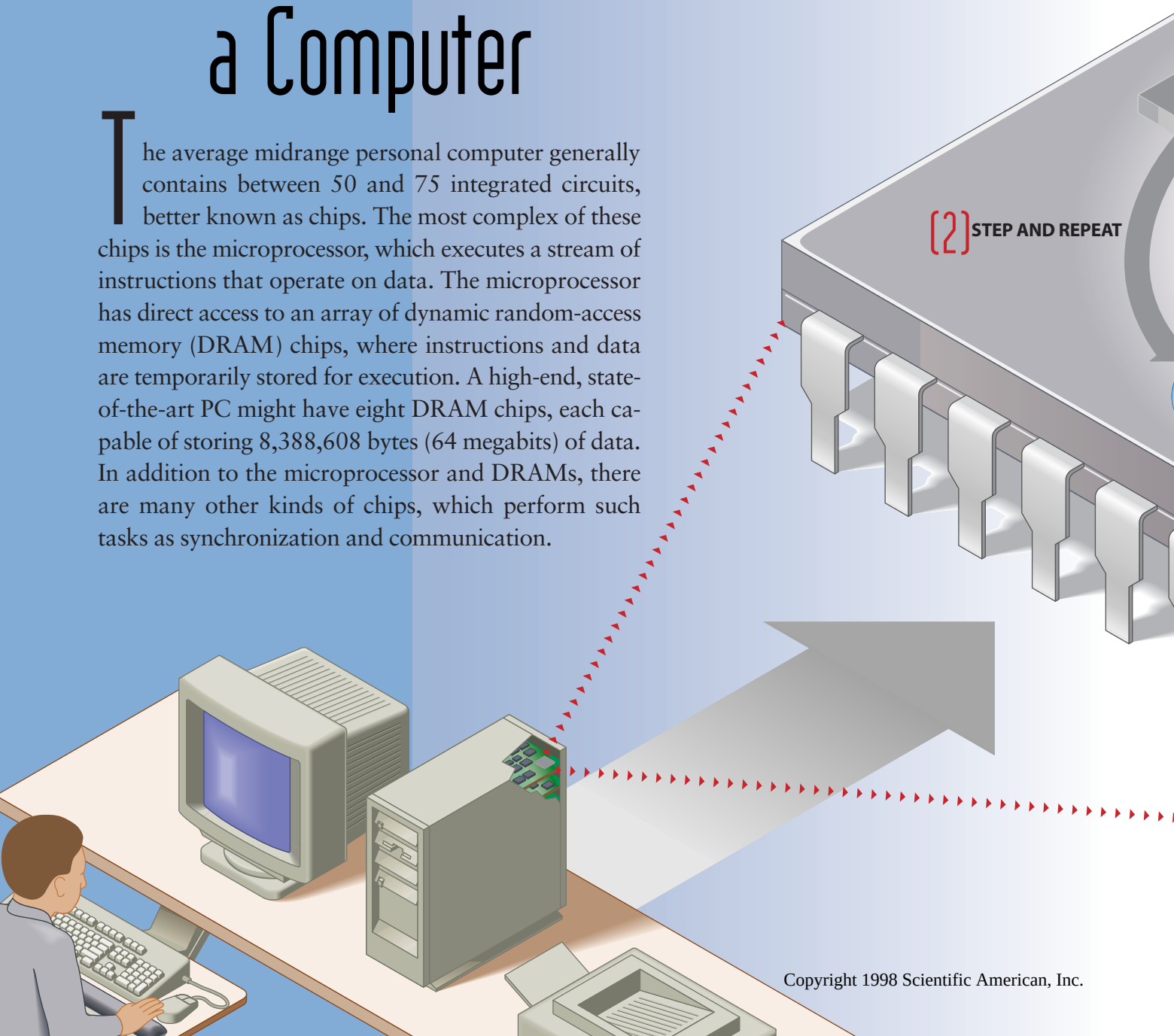
Frank H. Rockett is an electrical engineer and associate editor of the journal Electronics.

Copyright 1997 Scientific American, Inc.

Computers from Transistors

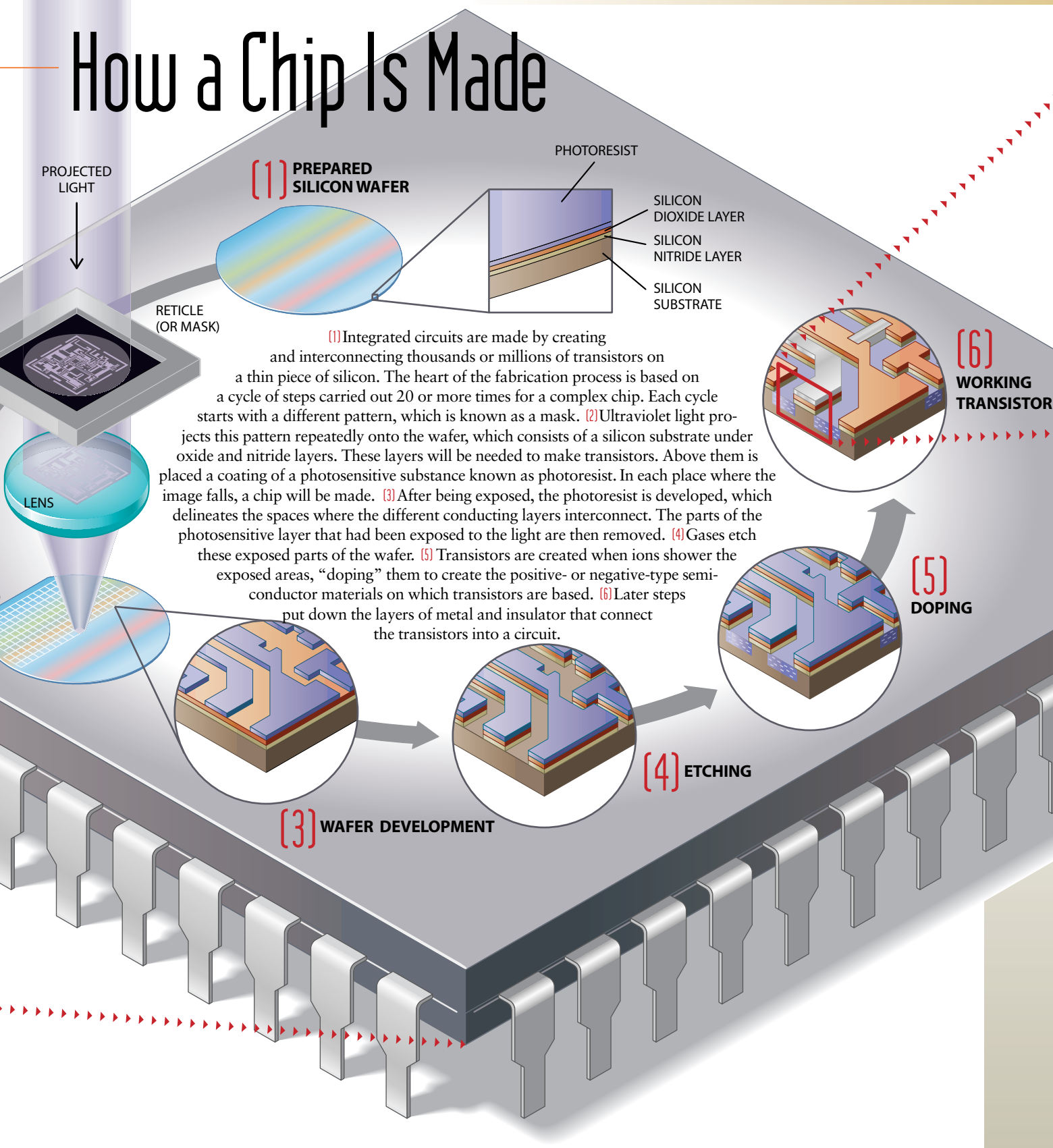
What's Inside a Computer

The average midrange personal computer generally contains between 50 and 75 integrated circuits, better known as chips. The most complex of these chips is the microprocessor, which executes a stream of instructions that operate on data. The microprocessor has direct access to an array of dynamic random-access memory (DRAM) chips, where instructions and data are temporarily stored for execution. A high-end, state-of-the-art PC might have eight DRAM chips, each capable of storing 8,388,608 bytes (64 megabits) of data. In addition to the microprocessor and DRAMs, there are many other kinds of chips, which perform such tasks as synchronization and communication.



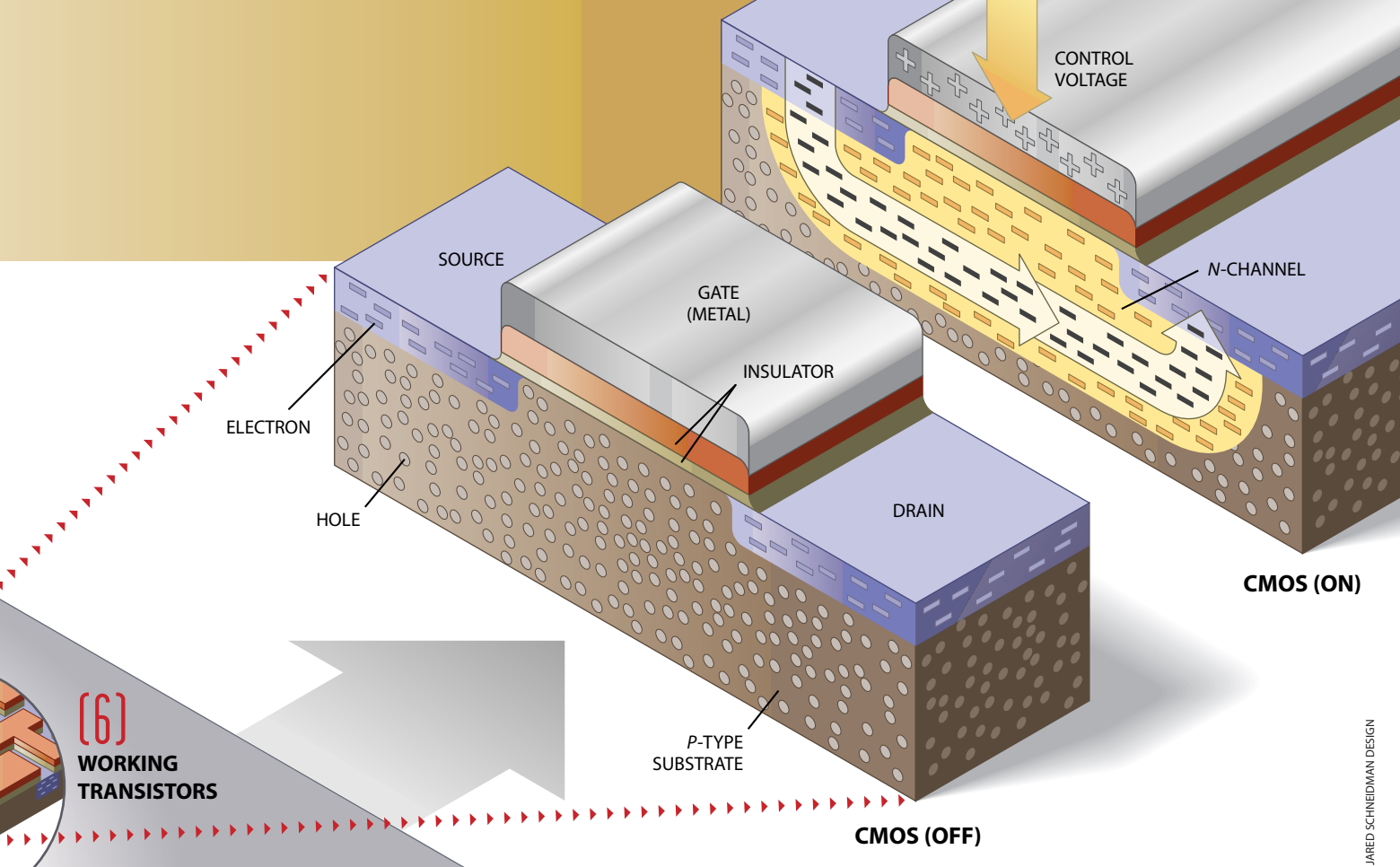
Copyright 1998 Scientific American, Inc.

How a Chip Is Made



(1) Integrated circuits are made by creating and interconnecting thousands or millions of transistors on a thin piece of silicon. The heart of the fabrication process is based on a cycle of steps carried out 20 or more times for a complex chip. Each cycle starts with a different pattern, which is known as a mask. (2) Ultraviolet light projects this pattern repeatedly onto the wafer, which consists of a silicon substrate under oxide and nitride layers. These layers will be needed to make transistors. Above them is placed a coating of a photosensitive substance known as photoresist. In each place where the image falls, a chip will be made. (3) After being exposed, the photoresist is developed, which delineates the spaces where the different conducting layers interconnect. The parts of the photosensitive layer that had been exposed to the light are then removed. (4) Gases etch these exposed parts of the wafer. (5) Transistors are created when ions shower the exposed areas, "doping" them to create the positive- or negative-type semiconductor materials on which transistors are based. (6) Later steps put down the layers of metal and insulator that connect the transistors into a circuit.

Copyright 1998 Scientific American, Inc.



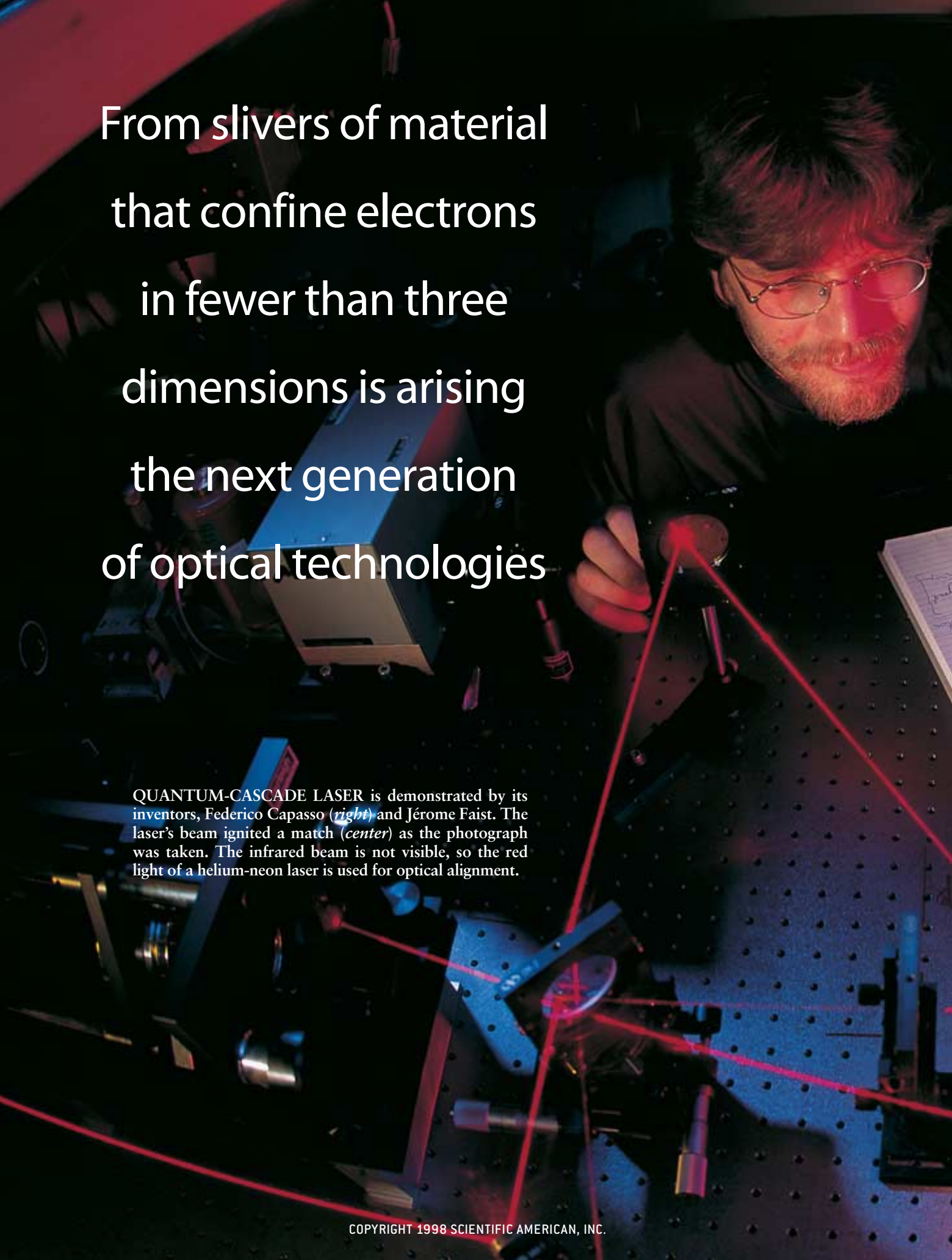
JARED SCHNEIDMAN DESIGN

How a CMOS Transistor Works

The transistors in an integrated circuit are of a type known as complementary metal oxide semiconductor (CMOS). They have two regions, the source and the drain, that have an abundance of electrons and are therefore referred to as *n* (for "negative") type. In between the source and drain is a *p*- ("positive") type region, with a surplus of electron deficiencies (called holes).

On top of the substrate, which is made of a silicon semiconductor material, is an insulating layer of silicon dioxide; on top of this oxide layer is a metal "gate." (Hence the name "metal oxide semiconductor.") When a positive voltage is applied to the metal gate, an electrical field is set up that penetrates through the insulator and into the substrate. This field attracts electrons toward the surface of the substrate, just below the insulator, allowing current to flow between the source and the drain.

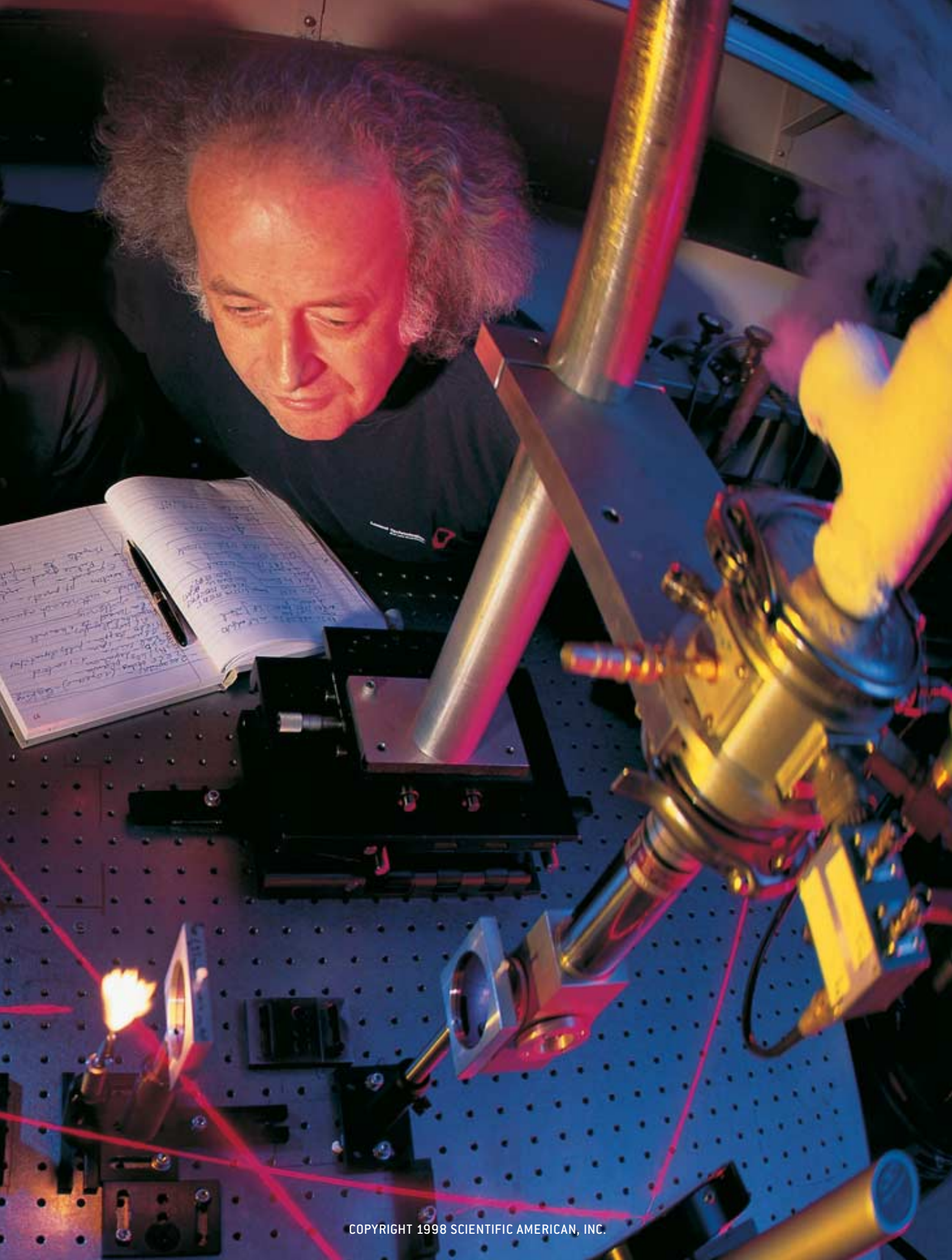
Copyright 1998 Scientific American, Inc.



From slivers of material that confine electrons in fewer than three dimensions is arising the next generation of optical technologies

QUANTUM-CASCADE LASER is demonstrated by its inventors, Federico Capasso (*right*) and Jérome Faist. The laser's beam ignited a match (*center*) as the photograph was taken. The infrared beam is not visible, so the red light of a helium-neon laser is used for optical alignment.

COPYRIGHT 1998 SCIENTIFIC AMERICAN, INC.



COPYRIGHT 1998 SCIENTIFIC AMERICAN, INC.

DIMINISHING DIMENSIONS

by Elizabeth Corcoran and Glenn Zorpette

In a tiny, cluttered room at Bell Laboratories, a division of Lucent Technologies in Murray Hill, N.J., researcher Jérome Faist is standing in front of an optical bench. In his right hand, near one of the lenses on the bench, he is holding a piece of paper torn from a desk calendar. In the middle of the bench, white puffs of water vapor pour from a cryostat, within which a revolutionary new type of laser known as a quantum cascade is being cooled with liquid helium.

With his left thumb, Faist taps a button on an instrument, boosting the voltage being applied to the semiconductor laser housed in the cryostat. Bright pinpoints of light on the piece of paper and wisps of smoke indicate that we have ignition. "If you need more convincing, you can put your finger in there," says a grinning Federico Capasso, with whom Faist invented the laser.

Burning paper with a laser is an old trick. But in this case there is a very new twist. The quantum-cascade laser is a speck of semiconductor material roughly the size of the capital letters on this page. Yet it is putting out 200 milliwatts at a wavelength of five microns, smack in the center of the middle-infrared region of the electromagnetic spectrum. Not only is the laser powerful, it is versatile as well: it can be tailored to emit light at essentially any frequency within a wide swath of the spectrum—something no other semiconductor laser can do.

Faist, a tall man with wire-rimmed spectacles and shaggy brown locks, is smiling broadly. "Just think," he says in his Swiss-French accent, "we can do much more clever things with this device than burn paper."

That is putting it mildly. Bell Laboratories's quantum-cascade laser is a dramatic confirmation that a new era in optoelectronics is under way. Lasers and similar devices will increasingly be built to exploit quantum effects—the peculiar, discrete behavior of subatomic particles, especially electrons, that have been confined to ultraminute realms in fewer than three dimensions.

Among the most promising applications are lasers, such as the quantum cascade. Capasso and Faist are now striving to build one that could operate continuously and at room temperature in the mid- or far-infrared part of the electromagnetic spectrum. Such a device could become the heart of spectroscopic instruments used to measure minute concentrations of airborne molecules—pollutants or contaminants, for instance.

The theory behind quantum devices has been known for decades. But in recent years the technologies that make such confinement possible by building infinitesimal structures out of individual atomic layers or molecules have been advancing at a remarkable pace. By controlling precisely the structure and composition of layers of materials tens of atoms or even just a few atoms thick, scientists are proving they can program the electronic characteristics they want into a compound. "It's like having your hands on the knobs of nature," says Mark A. Reed, head of the electrical engineering department at Yale University. Lucent's quantum-cascade laser, in particular, is an incredibly intricate layering of the semiconductors gallium indium arsenide and aluminum indium arsenide. Each layer is no more than 3.5 nanometers thick—several hundred thousandths of the thickness of a hair.

Confined within such a thin sheet of material, an electron takes on peculiar properties. In the macroscopic world we inhabit, the amount of energy in a system can vary continuously and smoothly. On an atomic scale, though, the energy of an electron orbiting, say, a proton in a hydrogen atom can be at only one of a number of well-defined, discrete levels. The electron need not be part of an atom to exhibit this quantum-energy effect; it is necessary only for the electron to be confined to a region whose dimensions measure anywhere from a few to a few hundred atoms. This characteristic size, known as the electron wavelength, approximates the hypothetical, indistinct cloud consisting of myriad points, each of which represents the probability of the electron occupying that position.

Thus, it makes sense to speak of an electron's wavelength in a semiconductor material—which is about 10 nanometers. Consider Lucent's quantum-cascade laser: electrons are free to occupy the slices of gallium indium arsenide. Partially confined to these semiconductor planes of only a few nanometers thick, the electrons begin to exhibit quantum behavior, such as having well-defined energy

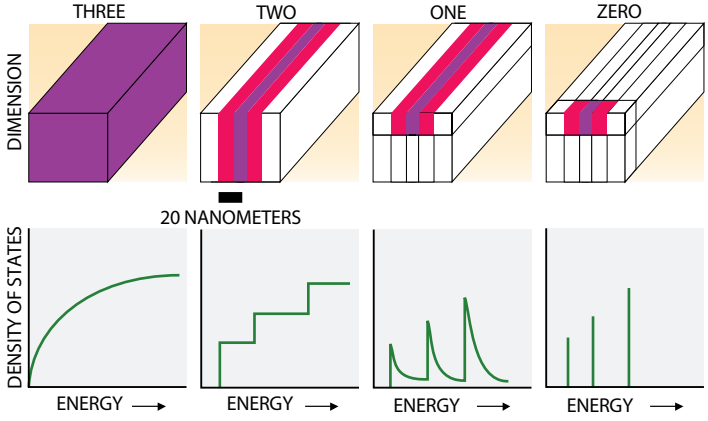
ELIZABETH CORCORAN is a former staff writer at Scientific American.

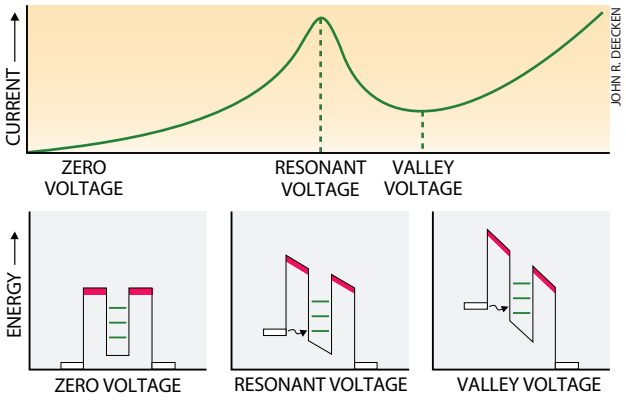
ROGER TULLY

COPYRIGHT 1998 SCIENTIFIC AMERICAN, INC.

Diminishing Dimensions

The dimensionality of a material can be reduced by sandwiching it between two layers of another material that has higher-energy electrons. This confinement changes the density of electron states, or specific energy levels, that will be filled by incoming electrons (*left*). The current conducted by a quantum-well device, shown by the green energy levels (*right*), peaks when the energy level of the incoming electrons matches, or is in resonance with, an energy level of the quantum well. At higher or lower voltages, little current leaks through the device.





levels. Through clever materials design, these electrons can be induced to jump from one energy level to another in an organized way, causing them to perform another useful trick—typically, emitting or detecting photons of light.

Wells, Wires and Dots

Quantum wells—ultrathin, quasi-two-dimensional planes—are just one of the three basic components of quantum devices. A narrow strip sliced from one of the planes is a one-dimensional quantum wire. Dicing up a one-dimensional wire yields zero-dimensional quantum dots. Reducing the number of dimensions in this manner forces electrons to behave in a more atomlike manner. By controlling the physical size and composition of the different semiconductors in a device, researchers can induce predictable changes in electron energy. In this way, scientists can literally pick, or tune, the electronic properties they want. In theory, the fewer the dimensions, the finer the tuning. Creating a zero-dimensional, or quantum, dot is analogous to custom-designing an atom. Like an atom, a quantum dot contains a certain amount of electrons. But whereas the electrons are held in an atom by their attraction to the nucleus, electrons in a quantum dot are physically trapped within barriers between semiconductor materials.

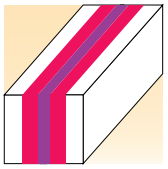
The only significant difference between an ordinary semiconductor laser and a quantum-well laser is in the relative size of each device's active region, where electrons and holes (electron deficiencies) recombine, neutralizing one another and causing a photon to be emitted. The quantum-well laser's active region is small enough for the energy levels of the electrons in the well to become quantized—that is, constricted to discrete values. This single difference, however, brings a major advantage: a quantum-well laser radiates light very efficiently, powered by much less current than a conventional semiconductor laser. As a result, semiconductor lasers that operate on the principle of quantum confinement dissipate far less excess heat. This feature, combined with the small physical size of the lasers, means that the devices can be packed tightly together to form arrays, are more reliable and can operate at higher temperatures.

What is true for quantum wells is even more so for quantum wires and dots—at least in theory. In practice, it has turned out to be quite a bit more difficult to exploit the advantages of the wires and dots than was expected a decade ago when the first such low-dimensional devices were built. Over the past few years, quantum-well semiconductor lasers have become commonplace. In fact, anyone who recently purchased a compact-disc player owns one. In contrast, quantum wires and dots are still in the laboratory. "Quantum wires and quantum dots are still miles from applications," Capasso notes. "But wells are already there."

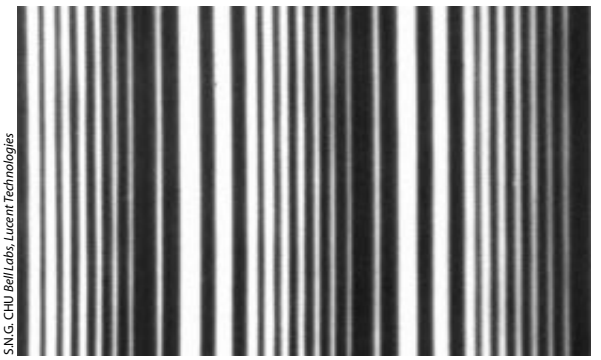
The difficulty of building useful quantum wires and dots has been sobering, after the intoxicating rush of advances in quantum devices in the 1980s and early 1990s. Researchers in those days envisioned two different classes of quantum devices: quantum optical devices, such as lasers and light detectors, and quantum electron devices, such as diodes and transistors. They even spoke enthusiastically of fundamentally different electron devices that, unlike today's binary "on-off" switches, would have three or more logic states. Functioning in parallel, these devices, it was hoped, would lead to more powerful forms of computer logic and become the building blocks of dramatically smaller and faster integrated circuits. There were also high hopes for so-called single-electron devices. These would include, for example, quantum dots that could contain so few electrons that the addition or removal of even a single electron would result in observable—and exploitable—effects. Using so few electrons, the devices could be switched on and off at blistering speeds and with very little power, investigators reasoned.

All these concepts were verified in laboratory demonstrations, but none resulted in anything close to a practical product. "The bottom line is that the sizes you need for useful semiconductors are just too small at room temperature," Reed says. "It's great science; it's just not a technology. That is not to say that there will never be some fantastic

Copyright 1998 Scientific American, Inc.



Two Dimensions: Quantum Well



S.N.G. CHU *Bell Labs, Lucent Technologies*

The white bands in this transmission electron micrograph are quantum wells consisting of gallium indium arsenide. The wells, which are sandwiched between barrier layers of aluminum indium arsenide, range in thickness from two atomic layers (0.5 nanometer) to 12 atomic layers (three nanometers). All the wells shown here are part of a single complete stage of a quantum-cascade laser, which comprises 25 such stages. When a voltage is applied to the device, electrons move from left to right, and each emits a photon as it tunnels between the two thickest quantum wells. Then the electron moves on to the next stage, to the right, where the process repeats, and another photon is emitted.

breakthrough that fundamentally changes things. But I'm pessimistic, frankly."

So, too, apparently, were IBM, Bell Communications Research (Bellcore) and Philips, all of which either abandoned quantum devices or severely curtailed their research programs in the mid-1990s. Nevertheless, in Japan, research into these devices continues unabated at large electronics firms and at many universities. A few U.S. and European academic institutions also continue to explore quantum-electron devices.

Yet even as work on these devices has stalled, enthusiasm is high for quantum optical devices, thanks to the quantum-well lasers, the quantum-cascade laser and a few other encouraging developments. Besides Lucent—which was recently spun off from AT&T—Philips, Thomson-CSF and Siemens have active research efforts. Many of those groups, including the one at Lucent's Bell Labs, hope to use such highly efficient, tiny quantum-well lasers to transmit data more efficiently and at higher rates through optical-fiber networks. One promising project at Lucent centers on a quantum-wire laser that promises lower-current operation. This laser would be desirable in a variety of applications, such as optical communications, because its low-current operation would enable the use of a smaller, less costly power supply.

And although experimentation with quantum electron devices and quantum dots may be down, it is certainly not out. Scientists at NTT Optoelectronics Laboratories in Japan, the University of California at Santa Barbara, the University of Southern California, Stanford University and the Paul Drude Institute in Berlin have begun investigating an intriguing new method of creating quantum dots, in which the infinitesimal devices sprout up as clumps on the surface of a semiconductor layer being grown with a technology known as molecular-beam epitaxy, the standard fabrication technique used to make quantum devices. And though hopes are fading for a commercially useful quantum-dot electron device in the near future, many researchers in academia are increasingly enthusiastic about quantum devices in which the electrons are contained by individual molecules, rather than semiconductor structures such as dots.

A Weird, Wonderful World

To build a lower-dimensional material deliberately, researchers must pay court to quantum mechanics. In any 3-D, bulk semiconductor, electrons take on a nearly continuous range of different energy states when additional energy is added to the material by applying voltage. As a result, researchers cannot tap a specific energy level; they must accept what they get.

Squeezing one side of a 3-D cube until it is no thicker than an electron's wavelength traps electrons in a 2-D plane. In two dimensions, the so-called density of electron states—the energy levels electrons can occupy—becomes quantized. Electrons jump from one energy level to another in a steplike fashion. After determining what layer thickness induces what energy level, workers can design the precise electronic characteristics of a material.

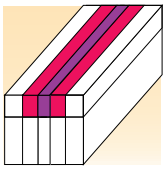
Electrons are not really confined by physical barriers; instead researchers must erect barriers of energy. Like water going downhill, electrons tend toward low-energy areas. So to trap electrons, investigators need only sandwich a material—typically a crystalline semiconductor filled with low-energy electrons—between two slices of semiconductor crystals with higher-energy electrons. Any electrons in the lower-energy slice will be confined, unable to traverse the interface, or barrier, between the two different semiconductor crystals if the barrier is sufficiently thick. The interface where the two crystals meet is known as a heterojunction. One of the few disappointing characteristics of silicon as a semiconductor material is that it does not emit light. So quantum-device builders use other, more exotic semiconductors, such as gallium arsenide and its many more complex compounds.

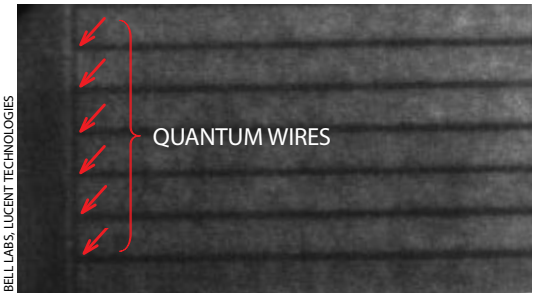
The energy of electrons in semiconductor crystals is described by band theory. When atoms are packed together to form a crystal, their energy levels merge to form bands of energy. Of particular interest are electrons in the valence band, because these electrons determine some of the material's properties, especially chemical ones. Valence electrons do not contribute to current flow, because they are fairly tightly held to atoms. To conduct electricity, electrons must be in a higher-energy band known as the conduction band. In metals, many of the electrons normally occupy this band, enabling them to conduct electric current.

A semiconductor, on the other hand, can be made to conduct substantial electric current by introducing impurities, called dopants, that deposit electrons into the conduction band. Electrons can also be introduced into the conduction band of a semiconductor by shining light into the crystal, which prods electrons from the valence band into the conduction band. The photocurrent generated in this way is exploited in all semiconductor light detectors, such as

Copyright 1998 Scientific American, Inc.

One Dimension: Quantum Wire





BELL LABS, LUCENT TECHNOLOGIES

The cleaved-edge overgrowth method creates quantum wires (*indicated by arrows*) by intersecting two seven-nanometer-wide quantum wells, which are essentially planar. The wells (and therefore the wires) are gallium arsenide; the barrier material outside the wells is aluminum gallium arsenide. Bell Laboratories researcher Loren Pfeiffer invented the cleaved-edge technique in 1991. An earlier method of creating quantum wires, pioneered at Bell Communications Research in the late 1980s, deposited the wire at the bottom of a V-shaped groove.

those in light-wave communications.

Alternatively, electrons can be injected into the conduction band by a voltage applied to electrical contacts at the surface of the crystal. Boosted to the conduction band, the electrons are able to take part in interesting phenomena, such as falling back to the valence band where they recombine with holes to produce photons of light.

The energy needed to propel an electron from the valence to the conduction band is the band-gap energy, which is simply the energy difference, typically measured in electron volts, between those two bands. Some semiconductors have higher- or lower-band-gap energies than others. Insulators, which require tremendous energy to push their valence electrons to the higher-energy bands, have the largest band gaps.

Scientists first began attempting to exploit these principles to build quantum electronics devices in the late 1960s. Thus, the era of quantum devices can be said to have begun 30 years ago, when Leo Esaki, Leroy L. Chang and Raphael Tsu of the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y., began trying to build structures that would trap electrons in dimensionally limited environments. "Confine an electron in two dimensions," Chang declared, "and it changes everything."

It was the invention of molecular-beam epitaxy (MBE) at Bell Labs by Alfred Y. Cho and John Arthur in the late 1960s that finally moved quantum research from the theoretical to the practical realm. At the heart of an MBE machine is an ultrahigh-vacuum chamber, which allows workers to deposit layers of atoms as thin as 0.2 nanometer on a heated semiconductor substrate. Attached to the vacuum chamber, like spokes on a hub, are three or four passages that lead to effusion cells. Elements such as gallium or aluminum are vaporized in these cells, then shot down the passages toward a substrate. By programming the shutters between the passages and the vacuum chamber, scientists can dictate the thickness of the layers deposited on the substrate, which is typically made of gallium arsenide or indium phosphide. Cho has likened the technique to "spray painting" atoms onto a substrate. The aim of both groups was to create a quantum well, which is made by depositing a very thin layer of lower-band-gap semiconductor between layers of higher-band-gap material.

At IBM, also using MBE, Esaki, Tsu and Chang began by alternating multiple layers of gallium arsenide with layers of aluminum gallium arsenide, a higher-band-gap compound. At about the same time, their counterparts at Bell Labs aimed to create a quantum well in a simpler way by sandwiching one thin, low-band-gap material between two higher-band-gap materials, thereby producing a quantum well. The idea was to trap electrons in the lower-band-gap semiconductor—gallium arsenide, for example, which has a band-gap energy of 1.5 electron volts. The electrons would be unable to cross the heterojunction barrier into the layers of aluminum gallium arsenide, which has a band gap of 3.1 electron volts. If the gallium arsenide layer—the actual quantum well—were just tens of atomic layers wide, quantum effects would be observed.

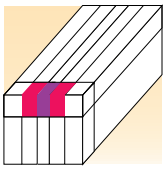
There was no arguing with the science, but at the time, it was ahead of the ability of the new MBE technology to exploit it. Efforts of both the IBM and AT&T groups bogged down in fabrication problems. For one, how do you lay down an even layer of material a few atoms deep? "We had to build a vacuum system ourselves" to deposit the ultrathin layers, says Chang, now dean of science at the Hong Kong University of Science and Technology. Equally troublesome was preventing contamination of the substrate, the material backing on which the thin layers would be deposited, in order to ensure a perfect meshing of the two different semiconductor crystal lattices at the heterojunction where they met.

In 1974 the researchers finally triumphed. The IBM team passed a current through a sequence of several quantum wells and observed peaks in the current as the voltage was increased. These peaks were caused by variations in the alignment of the energy levels in adjacent quantum wells and indicated that quantum confinement was occurring. At around the same time, Raymond Dingle, Arthur Gossard and William Wiegmann of Bell Labs built several isolated quantum wells, shone laser light on them and found that they absorbed different, but predicted, frequencies of light—an alternative indication of quantum confinement. Soon thereafter, Esaki and Chang of IBM built the first real quantum-well device—a resonant tunneling diode. As its name implies, the diode exploited tunneling, one of the most intriguing of quantum effects.

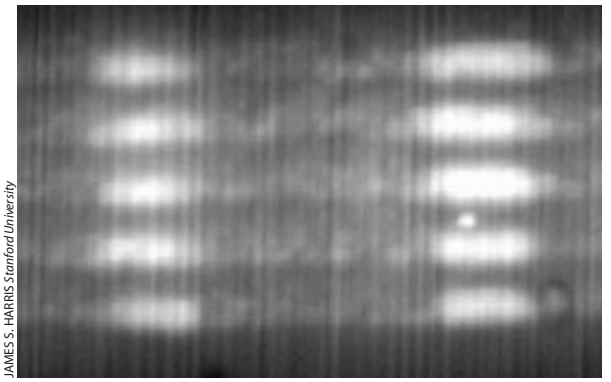
To understand tunneling, consider the classic quantum well described above. Typically, electrons are trapped between two high-band-gap semiconductors in the lower-band-gap, 2-D well between two relatively thick, high-band-gap semiconductor barriers. If the barriers are made sufficiently thin, a few nanometers, say, the laws of quantum mechanics indicate that an electron has a substantial probability of passing through—that is, tunneling through—the high-band-gap barriers.

Consider now an empty quantum well, surrounded by such ultrathin barriers. The whole structure, consisting of barriers and well, is sandwiched between electrically conductive contact layers. The trick is to apply just the right volt-

Copyright 1998 Scientific American, Inc.



Zero Dimension: Quantum Dot



JAMES S. HARRIS *Stanford University*

Three years ago researchers at Stanford University managed to produce multiple layers of quantum dots, in columns with the dots aligned vertically. Shown here are two columns, each with five dots. The top left dot is 18 nanometers wide; all the dots are about 4.5 nanometers high. The dots are indium arsenide; the surrounding barrier material is gallium arsenide. The ability to produce vertically aligned dots is considered an important step toward the integration of the dots into a useful device, such as a memory, in which each dot would be a memory element. A previous method, dating to the late 1980s, used lithographic techniques to create comparatively much larger dots.

age to the contact layers, so that the energy of the electrons entering the quantum well matches the energy level of the well itself. In this resonant tunneling phenomenon, many of the entering electrons will tunnel through the barriers, giving rise to a high current.

This passage of electrons can be exploited in various ways. Electrons can be caused to tunnel from one quantum well to another in a complex device, which has many quantum wells. (Bell Labs's quantum-cascade laser is such a device.) Alternatively, the tunneling can be the end result in itself, if the device is a diode or transistor, and the point is to have current flow.

Material Marvel

Although tunneling has so far proved a bust in the world of quantum-electron devices, its utility in optical devices has been ratified by the quantum-cascade laser, a material marvel. The QC laser, as it is known, is the first semiconductor laser that does not exploit recombinations of holes and electrons to produce photons and whose wavelength, therefore, is completely determined by an artificial structure—namely, the dimensions of a quantum well. It is also the most powerful mid-infrared semiconductor laser by far and the first that works at room temperature.

With respect to laser radiation, the mid- and far-infrared regions have been a barren land, where the few available semiconductor lasers are generally weak, cumbersome or constrained to narrow frequency bands. This lack of adequate mid- and far-infrared lasers has precluded the development of an entire class of spectroscopic devices capable of measuring minute concentrations of molecules—of pollutants or contaminants, for instance—in the air.

All such molecules absorb electromagnetic radiation at unique, characteristic and very specific frequencies. And many of these wavelengths, it so happens, are in the mid- and far-infrared range. So by tuning a laser, such as the quantum cascade, to the appropriate wavelength, shining the beam through air and measuring the amount of radiation absorbed, researchers could detect the presence and concentration of a certain molecule in the air. Environmentalists could use such a laser to monitor the emissions from a smokestack or from the tailpipe of an automobile. Semiconductor specialists could use it to check the cleanliness of a processing line, in which even a few stray molecules can render a chip useless. Law-enforcement and security officials could check for smuggled drugs or explosives. With sufficient power, such a laser might even be used on military jets, to "blind" or "fool" hostile heat-seeking surface-to-air missiles. (In fact, a significant part of the current funding for the quantum-cascade laser is coming from the Defense Advanced Research Projects Agency.)

The laser culminates more than a decade of tenacious pursuit by Lucent's Capasso, who at times was nearly alone in the conviction that it could be built. "I told Federico in the mid-1980s that it couldn't work," Yale's Reed says. "I'm happy to be proved wrong."

The fundamental requirement of a laser, regardless of its type, is to maintain a large number, or "population," of electrons in an excited state. The electrons are promoted to this excited state, which we'll call E_2, by applying energy from some external source. These electrons emit a photon of radiation when they drop down to a lower energy state, E_1. To achieve laser action, two conditions have to be satisfied. First, the higher energy level, E_2, must have a larger number of electrons than the lower one, E_1. This state, known as a population inversion, ensures that light is amplified rather than attenuated.

The second requirement for laser action is that the semiconductor material in which the photons are being generated must be placed between two partially transparent mirrors. This placement allows photons generated by electrons jumping from E_2 to E_1 to be reflected back and forth between the mirrors. Each time these photons traverse the material, they stimulate more electrons to jump from E_2 to E_1, emitting yet more photons, leading to laser action (hence the acronym: *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation). In a conventional semiconductor laser and also in the QC laser, the mirrors are built into the laser material. These perfectly smooth and reflecting facets are obtained by cleaving the semiconductor bar along crystalline planes.

Maintaining a population inversion demands that electrons be cleared away from the lower energy level, E_1. To do this requires yet another level, E_0, to which the electrons can be deposited after they have served their purpose. In the quantum-cascade laser, these three energy levels are engineered into a series of active regions, each consisting of three quantum wells. But there is not a one-to-one correspondence between the three energy levels and the three wells; rather the energy levels can be thought of as existing across all three wells. Some extremely intricate materials processing enables electrons to tunnel easily from one well to the next. It is this strong

Copyright 1998 Scientific American, Inc.

coupling that causes the various quantized energy levels to be, in effect, shared among the three quantum wells.

In operation, electrons pumped to a high-energy level tunnel into, and are captured by, the first two quantum wells. These electrons are at E_2. The electrons then drop from E_2 to E_1, emitting a photon in the process. The wavelength of the photon is determined by the energy difference between E_2 and E_1, as discovered by the Danish physicist Niels Bohr in 1913. Thus, Capasso and his colleagues can tune the wavelength of the emitted radiation merely by adjusting the width of the quantum wells to give the desired difference between E_2 and E_1. Using the same combination of materials, their laser spans the wavelength range from four to 13 microns.

The electrons then tunnel into the third quantum well, where they drop to E_0 before tunneling out of this last well and exiting the three-well complex. In order to maintain the population inversion, the electrons can persist at E_1 for only an extremely short period. This transience is guaranteed through yet more materials science wizardry: specifically, by engineering the difference in energy between E_1 and E_0. Capasso's group ensures that electrons do not linger in E_1 by setting this energy difference to be equal to that of a single phonon. In other words, to drop from E_1 to E_0, the electron need only emit one phonon. A phonon is a quantum of heat, much like a photon is of light. Put another way, a phonon is the smallest amount of energy that the electron can lose in dropping from one level to another.

Ingenious as they are, these features are not what makes the QC laser so unique. The laser's most novel characteristic is that it generates photons from not one of these three-quantum-well complexes but rather from 25 of them. The three-well complexes, which are known as active regions, are arranged in a series. Each successive active region is at a lower energy than the one before, so the active regions are like steps in a descending staircase. In between the active regions are injector/relaxation regions, which collect the electrons coming out of one active region and pass them on to the next, lower-energy one. All the active and injector/relaxation regions are engineered to allow electrons to move efficiently from the top of the staircase to the bottom. The end result is that a single electron passing through the laser emits not one photon but 25.

So far Capasso's group has achieved continuous emission only at cryogenic temperatures. Recently the laser set a record for optical power—200 milliwatts—at 80 kelvins, about the temperature of liquid nitrogen. The workers have also achieved, at room temperature, pulses that peak at 200 milliwatts. They aim to make their lasers work continuously at room temperature, a feat that could lead to all manner of portable, compact sensing devices. It is not an unrealistic goal, they say, given the development pattern that has occurred with other semiconductor devices. "There are no physics that forbid it," says Faist, who started working with Capasso in 1991. "It's definitely feasible."

Wiry Semiconductors

While some hail the QC laser as the latest manifestation of the utility of quantum wells, others wonder when the more exotic quantum structures—wires and dots—are going to start catching up. Physics suggests that if the two-

Electrons Traveling One by One

Current in conventional electronic devices is considered a kind of flowing river of electrons, in which the electrons are as uncountable as molecules of water. Reduce the dimensions of the material, and the energy of those electrons becomes quantized, or divided into discrete increments. Still, the precise number of electrons defies calculation.

Now, at the National Institute of Standards and Technology (NIST) in Boulder, Colo., researcher Mark Keller is building a system to make ultra-accurate measurements of capacitance, a form of electrical impedance, by precisely counting the number of electrons put on a capacitor. The heart of Keller's creation is a circuit that can count some 100 million electrons, give or take just one. This tally, along with a commensurate measurement of the voltage on the capacitor, will be used to determine capacitance with extreme accuracy. Thus, the capacitor will become a standard, useful to technological organizations for such applications as calibrating sensitive measuring equipment.

Keller's system is an expanded version of the electron turnstile invented in the late 1980s by researchers at Delft University in the Netherlands and at the Saclay Center for Nuclear Research in France. In those days, the Delft and Saclay workers were trying to build an electron counter that could be used as a standard for current, rather than capacitance. Ultra-accurate electron counts are, in theory at least, useful for setting a standard for either quantity.

The central part of the electron turnstile was an aluminum electrode about one micron long and coated with aluminum oxide. This bar was known as an island because it was isolated on each side by a nonconductive junction connected to a metallic electrode, or "arm." When a large voltage was applied across this device, between the two arms, it behaved like a conventional resistor. But at temperatures of about one kelvin and voltages of a few tenths of a millivolt, the resistance increased dramatically. No current could flow through the device because the electrons had insufficient energy to get past the junctions and onto the island. Increasing the voltage to about one millivolt, however, gave electrons just enough energy to tunnel from the arm, across the junction and to the island.

To control the flow of electrons, the voltage was applied to the island through a capacitor (not a standard capacitor but a high-quality but otherwise ordinary capacitor). As the capacitor charged and discharged, the voltage increased and decreased, forcing one electron onto, and then off, the central island. An alternating current, fed to the capacitor, controlled the charging and discharging; thus, the frequency of the alternating current determined the rate at which individual electrons passed through the island.

The concept was elegant, but the implementation fell short of the mark. The Delft and Saclay workers were able to get accuracies of one electron in about 1,000. For them to top the existing standard for current, precision of better than one in 10 million would have been necessary. The problem with this single-island setup was that occasionally two electrons, or none at all, would pass through the island during a cycle of the alternating current. Moreover, current flow through the turnstile, at about a picoampere (0.000000000001 ampere) was too low for useful metrology.

Copyright 1998 Scientific American, Inc.

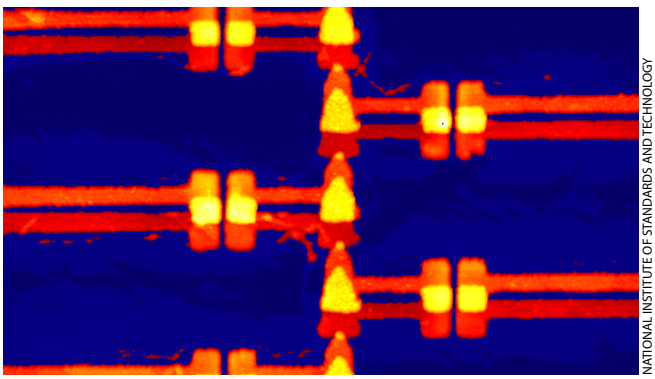
dimensional realm is promising, then one or zero dimensions is even better. Electrons scatter less and attain higher mobilities when traveling through a quantum wire than through a plane. What this means is that lower-dimensional lasers could be powered by far less current. Such lasers would also have a lower lasing threshold, which means that lower populations of free electrons and holes would be necessary to get them to emit laser radiation. This characteristic in turn would mean that the lasers could be modulated at higher frequencies and therefore transmit information at a higher rate.

That kind of incentive is keeping researchers interested in quantum-wire lasers despite daunting challenges. To make the wires, they must wall up four sides of a low-band-gap material with higher-band-gap barriers thin enough to let electrons tunnel through on command—about an electron wavelength thick. Exercising the precise control needed to make such vertical walls is tricky, to say the least.

Several techniques have been developed. A quantum well already has two barriers, and one method simply etches two more barriers chemically, using a lithographic technique. The wire, then, exists between these barriers and above and below the well's heterojunctions. The etched barriers, however, tend to confine poorly in comparison with the heterojunctions, and therefore the cross section of the wire turns out to be a rather long oval, roughly 50 by 10 nanometers. Another technique, pioneered at Bellcore in the late 1980s, deposits the wire using MBE techniques at the bottom of a V-shaped groove. These wires also suffer from some of the drawbacks of the ones created through lithography.

Currently, the leading technique for creating symmetric quantum wires is the cleaved-edge overgrowth method, first demonstrated at Bell Labs by researcher Loren Pfeiffer in 1991. The technique is now in use at Lucent, the Research Center for Advanced Science and Technology at the University of Tokyo and the Walter Schottky Institute in Garching, Germany. The method creates two quantum wells that intersect perpendicularly in a quantum wire. The first well, a seven-nanometer-thick layer of gallium arsenide sandwiched between layers of aluminum gallium arsenide, is grown using conventional MBE. Researchers rotate the sample 90 degrees and scratch it to initiate the cleft. The sample is then broken cleanly at the scratch to create an atomically sharp, perfect edge. Then MBE resumes putting down new layers—this time on top of the cleaved edge. Thus, the technique creates two perpendicular planes of gallium arsenide, which intersect in a quantum wire with a cross section seven nanometers wide.

In 1993 Pfeiffer and his colleagues demonstrated that a quantum-wire laser has an unusual property: it emits photons that arise from the recombination of excitons, which are bound electron-hole pairs, analogous to the binding between an electron and a proton in a hydrogen atom. In a conventional semiconductor laser or even in a quantum-well laser, on the other hand, the densely packed excitons interact, disrupting the relatively weak pairing between electron and hole. The resulting electron-hole plasma still generates photons, but from the mutual annihilation of free electrons and free holes, rather than from the recombination of electrons and holes already paired together in excitons.



NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

JUNCTIONS are the small, bright dots where one island touches the one above it.

In the early 1990s researcher John Martinis of NIST picked up on the work but did so in hopes of producing a standard for capacitance rather than for current. Shortly before, the Saclay researchers had expanded their electron turnstile into an electron "pump," with two islands and three junctions. Martinis then built a pump with four islands separated by five nonconductive junctions. The numbers were chosen because theoretical calculations based on the physics of tunneling had indicated that they would suffice to achieve an accuracy of one part in 100 million, the figure needed to create a competitive standard for capacitance.

In these pumps, alternating current fed to a capacitor still controlled the voltage applied to each island. But these currents were synchronized so that the voltages were applied sequentially down the chain of islands, in effect dragging a single electron through the chain from island to island. When incorporated into a capacitance standard, sometime in the near future, the circuit will be arranged so that the electrons that go through the chain of islands will wind up on a standard capacitor. Thus, the number of cycles of the alternating current will determine the number of electrons on that standard capacitor.

By offering electrical control over each of the junctions, the electron pump turned out to be considerably more accurate than the electron turnstile. Yet the sought-after accuracy of one part in 100 million was still out of reach. In a paper published in 1994, Martinis reported that he had achieved accuracy of five parts in 10 million. Researchers were unsure why the accuracy fell short of the theoretical prediction.

Along came Mark Keller, who joined NIST as a postdoctoral employee in 1995. Keller extended the electron pump to seven islands and, just recently, achieved the desired accuracy of one part in 100 million. He is now working to turn the circuit into a practical capacitance standard, using a special capacitor developed by NIST's Neil Zimmerman. The capacitor does not "leak" charges and is unusually insensitive to frequency.

With the metrological goal achieved, Keller and Martinis have turned their attention to the nagging mismatch between experiment and theory. They believe they have identified an important source of error, unacknowledged in the original theory. The error, they suspect, is caused by electromagnetic energy from outside the pump, which gets in and causes electrons to go through junctions when they should not. The two researchers are now conducting painstaking measurements to test the idea. Sometimes, it would appear, good science comes out of technology, rather than the other way around. *—E.C. and G.Z.*

Copyright 1998 Scientific American, Inc.

This seemingly small difference in physics leads to a major difference in the way that the lasers radiate. As the intensity of an ordinary semiconductor laser is increased (say, by boosting the current) the energy of the photon emissions from a free-electron-hole plasma is reduced. This phenomenon, called band-gap renormalization, causes the laser's emission frequency to shift downward, which could inhibit the performance if the laser is being used for spectroscopy or to transmit information. In the intense confinement of a wire or dot, on the other hand, the excitons do not fragment, so the frequency remains stable when input current, and therefore output power, is increased.

Earlier this year Pfeiffer and his colleague Joel Hasen found that at low temperatures, their quantum wires metamorphose in such a way that exciton emission comes not from a uniform wire but from a series of dozens of quantum dots spread out along the 30-micron length of the wire. At these low temperatures, the unevenness of the wire's width has an interesting effect on its quantum behavior. To understand this effect requires a little background on MBE. Because of limitations in even the best MBE systems, a uniform quantum wire cannot be made, say, 24 atomic layers wide for its entire length. In some places it may be 23, and in others, 25 (these differences are known as monolayer fluctuations).

At low temperatures, excitons are less able to penetrate the narrower, higher-energy parts of the wire; thus, these narrow areas become de facto barriers along the wire. They wall off sections of the wire, creating a string of dots. It is too soon to say whether this phenomenon will give rise to any practical applications. But it has already prompted Pfeiffer and Hasen to make plans to study the differences between the radiative properties of wires and of dots in lasers. "This is the first quantum system where you can change the temperature and go between two different regimes: wires and dots," Hasen declares.

The Zero Zone

In fact, the quantum dot, the ultimate in confinement, is still the subject of intensive research, particularly in university laboratories in North America, Japan and Europe. Quantum dots have been called artificial atoms, in spite of the fact that they generally consist of

Physics suggests that if the two-dimensional realm is promising, then one or zero dimensions is even better.

thousands or hundreds of thousands of atoms. Confined in a dot, or box, electrons should occupy discrete energy levels. It should be possible, therefore, to dial up precise energy levels by adjusting the construction of the quantum box and by varying the applied voltage. In the 1980s and early 1990s researchers created dots by using lithographic techniques similar to those used to make integrated circuits. Success was sporadic and limited to relatively small numbers of dots, which were essentially useless as lasers.

The picture began to change a couple of years ago with the invention of so-called self-assembly techniques. Researchers had noticed that oftentimes, clumps would form spontaneously on the surface of extraordinarily thin layers of certain materials grown with MBE. Initially considered something of an annoyance, the phenomenon was actually rather intriguing.

Suppose a single monolayer of indium arsenide is grown on a substrate of gallium arsenide. This single monolayer perfectly and evenly covers the gallium arsenide. As more indium atoms are added, however, the perfect coverage ceases. Specifically, when enough atoms have been laid down so that the average coverage is between about 1.6 and 1.8 monolayers, the clumping begins. "It's really amazing," says James S. Harris, professor of electrical engineering at Stanford. "It is incredible that it occurs in such a narrow window." By the time enough material has been laid down for three even monolayers, what has formed instead is an aggregation of an enormous number of clumps, each five or six monolayers high, separated by much shallower regions.

Scientists soon realized that these clumps, resembling tiny disks four to five nanometers high and 12.5 nanometers in diameter, could function as quantum dots. Though somewhat larger than that would be ideal, the dots are easily and quickly made by the millions. And in 1994 Harris succeeded in producing multiple layers of the dots in a single crystal. The dots were created in such a way that those in one layer were all perfectly aligned with the ones above and below [*see illustration on page 29*]. The achievement was an important step toward the integration of many dots into a useful device—a memory device, for example, in which each dot is a memory element.

Although electron devices such as memories are a distant possibility, optical devices such as lasers have already been demonstrated. In 1995 Dieter Bimberg of the Technical University of Berlin coaxed laser radiation from an array of perhaps a million of the layered dots for the first time. Since then, several groups, including ones at the National Research Council of Canada in Ottawa, at a Fujitsu laboratory in Japan and at the Ioffe Institute in St. Petersburg, Russia, have also managed to draw laser radiation from the dots. Harris contends that the radiation, which is in the near-infrared, comes from the dots in these arrays and not from the underlying substrate, which is actually a quantum well. Other researchers are less convinced. Harris adds that the dot arrays have impressive characteristics as lasers but are not now in the same league with the best quantum-well lasers.

Other research teams working on self-assembled dots include ones at the University of California at Santa Barbara, at the French telecommunications research laboratory CNET and at two German research organizations, the Max Planck Institute in Stuttgart and the Technical University of Munich.

Meanwhile, at the IBM Thomas J. Watson Research Center, Sandip Tiwari is trying to use silicon to build a memory system based on quantum dots. Tiwari starts with a very thin layer of silicon dioxide, on which he seeds silicon quantum dots, exploiting a phenomenon similar in some respects to self-assembly. Tiwari is studying the characteristics of single and multiple dots—for example, the change in electric field as an electron is added or removed from a dot containing five to 10 electrons.

The Molecule as Dot

In projects similar to Tiwari's, several research groups recently realized one of the most sought-after goals of quantum electronics: a memory system in which a bit is stored by inserting or re-

Copyright 1998 Scientific American, Inc.

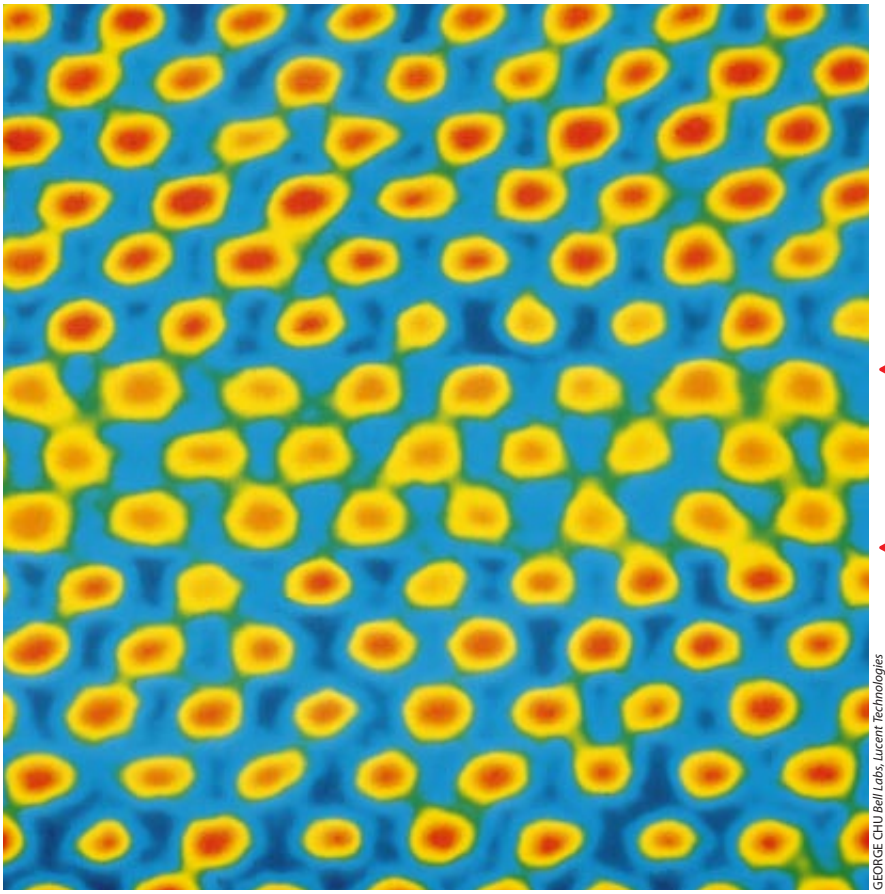
moving an electron from an otherwise empty quantum dot. In this system, each dot has an associated transistorlike device, which also operates with individual electrons and "reads" the dot by detecting the presence (or absence) of an electron from its electric field. The dots used in these experiments are produced using lithography rather than self-assembly, because of the difficulty of linking self-assembled dots to the transistorlike devices. Operating as they do with just one electron, and in zero dimensions, these devices are in effect quantized in charge as well as in space.

Researchers at such institutions as Harvard University, S.U.N.Y.–Stony Brook, Notre Dame, Cambridge, Hitachi, Fujitsu and Japan's Electrotechnical Laboratory have built single-electron dots, transistors or complete, working memory cells. Unfortunately, the experiments have produced small numbers of devices, some of which can be operated only at extraordinarily low temperatures. Most inauspiciously, no promising methods for connecting millions of the devices, as would be required for a practical unit, have emerged.

Just as hopes for practical, single-electron quantum dots are fading in some circles, they are rising for an alternative approach: using molecules, instead of synthesized quantum dots, for the confinement of single electrons. "There has been a realization over the past year that if zero-dimension, single-electron quantum devices do happen in a technologically useful way, they're going to happen in the molecular area, not in the semiconductor area," says Yale's Reed.

Over the past couple of years, researchers have managed to measure the characteristics of individual molecules. At the University of South Carolina, professor James M. Tour was part of a group that began a few years ago with wires created by linking benzene-based molecules into a sort of chain. The researchers also produced the molecular equivalent of alligator clips, which are affixed to the ends of the benzene-based "wires." The clips consist of thiol molecules (linked sulfur and hydrogen atoms), which let them connect the wires to metal substrates or other molecules.

Last year Tour was part of another team that connected one of the wires to a gold substrate. To measure the conductivity of the wire, the researchers contacted one end of the wire with the tip of a scanning-tunneling microscope (STM). To ensure that they were measuring the conductivity of the wire alone, the researchers had inserted the wire into a relatively dense "thicket" of nonconductive alkanethiol molecules. They found that the wires, which were about 2.5 nanometers long by 0.28 nanometer wide, had high conductivity relative to that of other molecules that had been probed and tested in this way.



GEORGE CHU *Bell Labs, Lucent Technologies*

QUANTUM WELL is three atomic layers of gallium indium arsenide (*horizontal strip indicated by two arrows*) between layers of indium phosphide. Blue areas in this false-colored transmission electron micrograph show the positions in the crystalline lattice of atoms, which are separated by a mere 0.34 nanometer (340 trillionths of a meter).

More recently, a collaboration between Yale and the University of South Carolina measured the electrical conductivity of a single benzene-based molecule situated between two metallic contacts. Unfortunately, the electrical resistance of the setup was mainly in the thiol alligator clips, between the molecule and the metallic contacts. So they wound up measuring mainly the resistance of the alligator clips. To get around the problem, they are working on more conductive clips. Still, Tour points out that the researchers succeeded in verifying that they were able to put only one electron at a time into the molecule. Thus, the device is the molecular embodiment of a single-electron quantum dot. "We got reasonable current, on the order of tenths of microamps, one electron at a time," Tour notes proudly.

Where is all this work headed? Ideally, to a molecule that acts like a transistor. Such an infinitesimal device would have to have at least three contact points, or terminals, if current flow between two of the terminals were to be controlled by another. Although three- and four-terminal molecules have been simulated on a computer and even produced in the laboratory, the challenges of testing them are prohibitive. "Three-terminal molecules are so small, you can't bring the scanning-tunneling-microscope tips, which are macroscopic, close together enough to contact all three terminals," Tour says.

It is too soon to say whether quantum electronics will make much progress along this particular route. But it is clear that as miniaturization lets optoelectronics and electronics delve deeper into the strange, beautiful quantum world, there will be intriguing and splendid artifacts. SA

Copyright 1998 Scientific American, Inc.

The insulated gate bipolar transistor is transforming the field of power electronics

HOW THE SUPER-TRANSISTOR WORKS

by B. Jayant Baliga

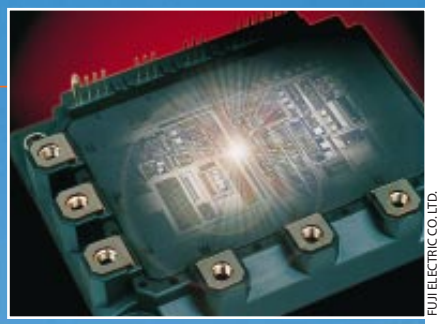
Although it is rarely acknowledged, not one but two distinct electronic revolutions were set in motion by the invention of the transistor 50 years ago at Bell Telephone Laboratories. The better known of the two has as its hallmark the trend toward miniaturization. This revolution was fundamentally transformed in the late 1950s, when Robert N. Noyce and Jack Kilby separately invented the integrated circuit, in which multiple transistors are fabricated within a single chip made up of layers of a semiconductor material. Years of this miniaturization trend have led to fingernail-size slivers of silicon containing millions of transistors, each measuring a few microns and consuming perhaps a millionth of a watt in operation.

The other, less well known, revolution is characterized by essentially the opposite trend: larger and larger transistors capable of handling greater amounts of electrical power. In this comparatively obscure, Brobdingnagian semiconductor world, the fundamental, transformative event occurred only a few years ago. And the golden era is just getting under way.

The seminal development in this field, known as power electronics, was the invention of a new kind of transistor, the insulated gate bipolar transistor (IGBT). These semiconductor devices, which are about the size of a postage stamp, can be grouped together to switch up to 1,000 amperes of electric current at voltages up to several thousand volts. Most important, IGBTs can switch these currents at extraordinarily fast speeds, making them far superior in every way to their predecessors.

Already IGBTs are being used as a kind of switch to control the power flowing in many different kinds of appliances, components and systems. In many of these items, groups of IGBTs are con-

Copyright 1997 Scientific American, Inc.

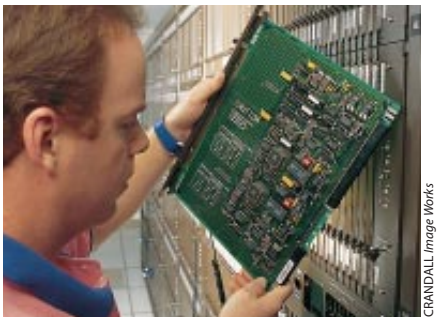


FUJI ELECTRIC CO. LTD.

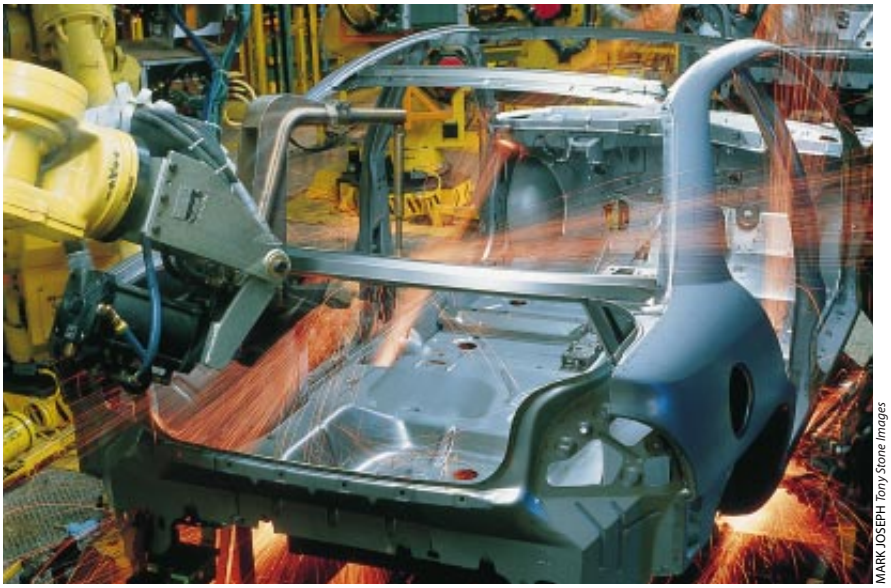
KIM KULISH *Sygma*

TONY STONE IMAGES

BLACK & DECKER



CRANDALL *Image Works*



MARK JOSEPH *Tony Stone Images*

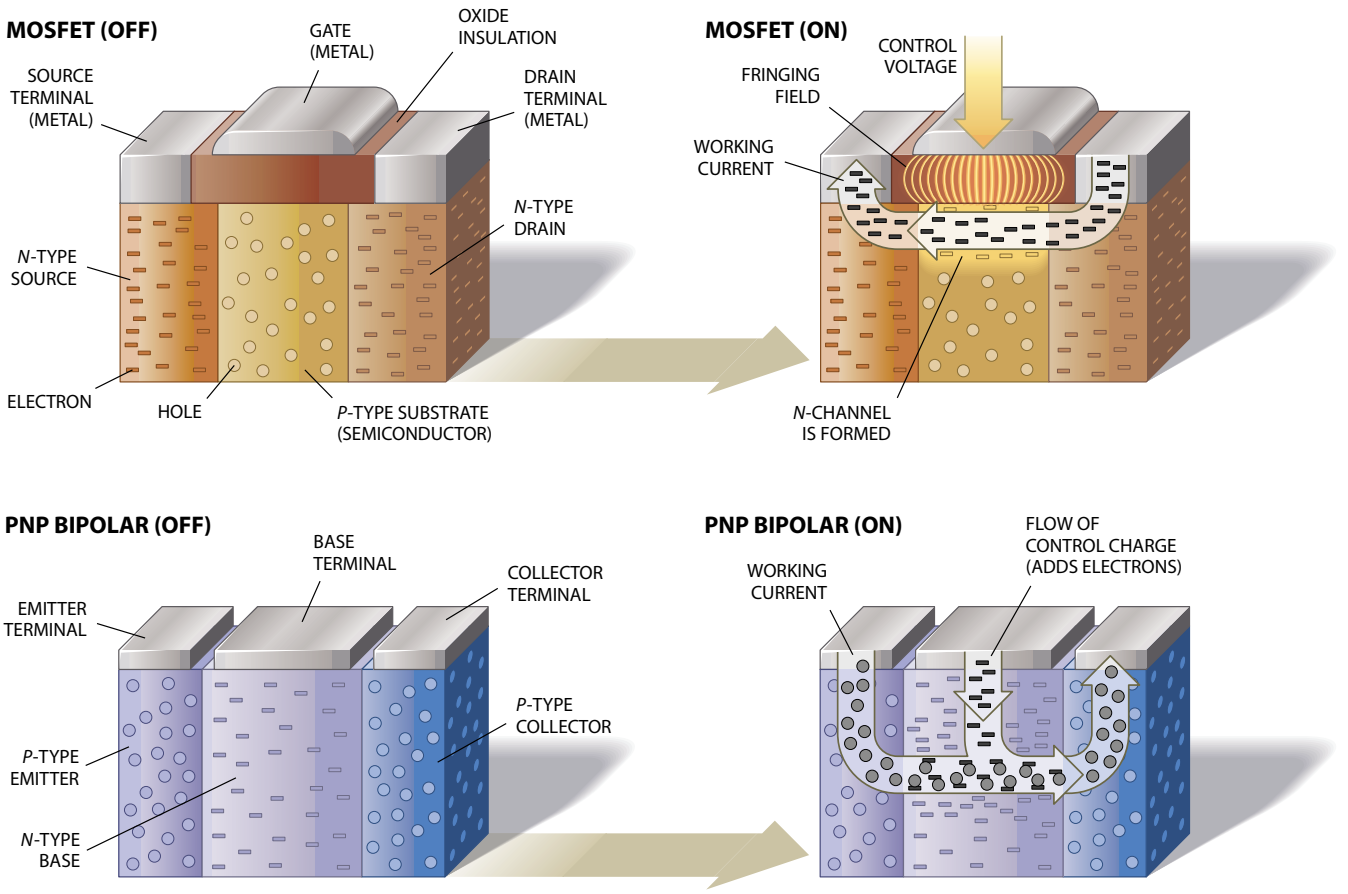
APPLICATIONS of the insulated gate bipolar transistor (IGBT) encompass a diverse group: steam irons, telephone-system central office switches, electric cars and high-speed trains. IGBTs are particularly attractive in factory automation, because precise movement of robotic arms demands superior motor controls. The device itself consists of a sliver of silicon encased in plastic (*upper left inset photograph*). The transistor's capabilities are impressive: IGBTs are available that can switch 1,000 amperes.

nected together to control the power applied to electric motors.

Electric-motor controls are a major business, with applications in both industry and the home. Factories, for example, generally rely heavily on motor-driven machinery, equipment or robots. Electrically powered streetcars and trains, too, need motor controls. The motors in Japan's famous Shinkansen bullet trains, for example, are now controlled by IGBTs. And the average household in a developed country has been estimated to have over 40 electric motors, in appliances such as blenders, power tools, washers and dryers and in the compressors of refrigerators, freezers and air conditioners. Essentially all electric cars built within the past few years also rely heavily on IGBTs.

The speed and power of most modern alternating-current motors, whether the motor is in a blender or a bullet train, are varied by altering the frequency and amplitude of the sine wave that is applied to the motor's windings. With this type of control system, which is known as an adjustable-speed drive, the motor's rotor turns with the same frequency as the sine wave. Groups of IGBTs can be used to create this sine wave by putting out pulses of precisely controlled duration and amplitude. Because IGBTs can be switched on and off so rapidly, they can produce a comparatively smooth sine wave. This smooth-

Copyright 1997 Scientific American, Inc.



MOSFET AND BIPOLAR TRANSISTORS are combined to create an IGBT, a rugged, high-power device. In the metal oxide semiconductor field-effect transistor, or MOSFET, current flow is enabled by the application of a voltage to a metal gate. The voltage sets up an electrical field that repels positively charged electron deficiencies, known as holes, away from the gate. At the same time, it attracts electrons, forming between the source and the drain a so-called *n*-channel, through which a working current flows. In the *p-n-p* bipolar transistor, a relatively small control current adds electrons to the base, attracting holes from the emitter. These holes flow from the emitter to the collector, constituting a relatively large working current. In the IGBT, a con-

ness in turn keeps the motor from generating excessive harmonics, which are stray sine waves with frequencies that are higher by a factor of two, three, four and so on. Harmonics create heat, waste energy and can damage the motor or other equipment on the circuit.

Before IGBTs, the motors used in, for example, heating, ventilating and air-conditioning (HVAC) units were typically run at constant speed and merely turned on and off at different intervals to accommodate changes in ambient temperature. Efficiency under slack loads was poor. Adjustable-speed drives based on IGBTs offer far superior efficiency, which has been estimated to save millions of barrels of oil every day, which also reduces pollution. These efficient HVAC controls have already been widely adapted in Japan and are increasingly popular in Europe and in the U.S.

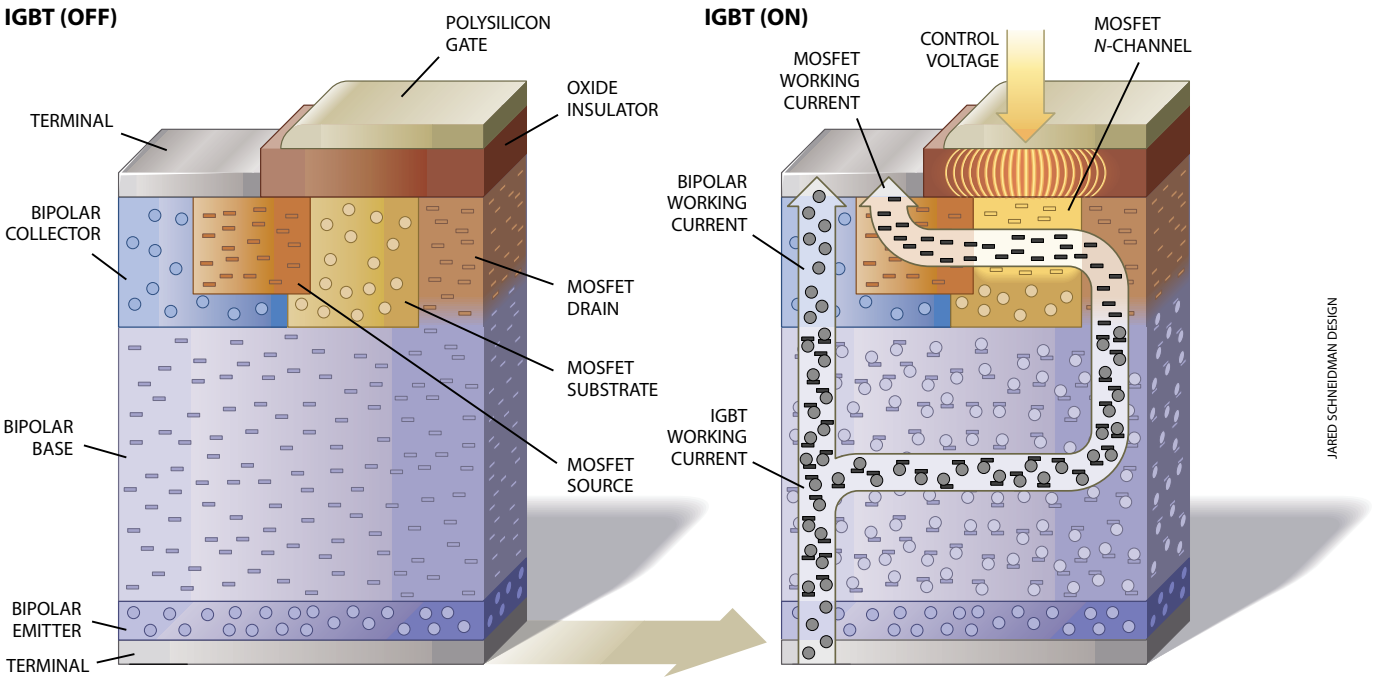
Another advantage of IGBTs stems from their switching speeds: they are so fast that the pulses they generate can easily have a frequency above the range of human hearing. Thus, IGBTs can be used to build completely silent compressors for air conditioners, refrigerators and the like. That hum that comes from most compressors is caused by slower power-electronics devices, which can be switched on and off only at frequencies within the range of hearing.

IGBTs can do a lot more than control motors. Some companies are now using them in the latest laptop-computer displays, to turn picture elements on and off. Telephone equipment manufacturers are incorporating them into central office switches, to route signals by connecting different circuits and also to activate the circuit that sends the signal that rings a telephone. One company has even used IGBTs to produce an advanced defibrillator, a lifesaving device that delivers an electric shock to restart the heart of a victim of cardiac arrest. IGBTs are also being used in the ballasts of fluorescent and arc-discharge lights, to regulate the power that surges through these gas-filled tubes, breaking down the gas's electrical resistance and causing it to emit electromagnetic radiation.

All told, power-electronics devices, including IGBTs, control an estimated 50 to 60 percent of the electrical power generated in industrial countries. Moreover, this percentage is growing, thanks mostly to the success of IGBTs.

As these devices begin dominating major parts of power electronics, they are finally uniting the two electronic revolutions that began half a century ago. IGBTs can use an electric current of just a few thousandths of an ampere to control flows of, say, 100 amperes at 1,500 volts. And their ability to be controlled by such minute currents enables IGBTs to be fabricated on the same semiconductor chip with the circuits that permit the IGBT to be controlled by microprocessors. To draw an analogy to physiology, if the microprocessor and its associated memory chips are like the brain,

Copyright 1997 Scientific American, Inc.



trol voltage is applied to a MOSFET. It establishes a working current, which in turn is applied—as a control current—to the base of a *p-n-p* bipolar. This control current enables a larger working current to flow in the bipolar. Because of the arrangement of its components, the IGBT's working current is actually the combined working currents of both the MOSFET and the bipolar. This ingenious configuration enables the devices to have a power gain—the ratio of the working current and voltage to the control current and voltage—of about 10 million. Such gain enables the devices to connect to microelectronic circuits.

IGBTs can be thought of as the muscles. Ñever before have brains and brawn been so intimately connected.

Fluorescent Lights to Bullet Trains

A deeper understanding of the ascendancy of IGBTs requires some perspective on power semiconductors. The market might be broken down into three voltage categories: low, medium and high [*see illustration on page 39*]. The first, comprising applications involving less than 100 volts or so, includes automotive electrical systems, certain power-supply circuits used in personal computers, and audio-power amplifiers, such as those used in stereo high-fidelity systems. This segment of the market is dominated by a kind of device known as a metal oxide semiconductor field-effect transistor (MOSFET), which might be thought of as today's ordinary, garden-variety transistor.

The middle category of voltages is a wide one, ranging from 200 to about 1,500 volts. This is the province of the IGBT. This category, moreover, can be subdivided into two others. At the lower end, such devices as laptop-computer displays, telecommunications switches and lighting ballasts all generally require devices capable of handling between 200 and 500 volts. Current flows are relatively small (less than 10 amperes), so there is a strong thrust toward putting the power-switching devices and the microelectronics that control them on the same chip.

In the higher end of this middle range, typical applications include motor controls and robotics, which demand devices that can handle between 500 and 1,500 volts. IGBTs are especially attractive for robotics because the precise movement of platforms and arms can be accomplished only with superior motor controls. An early implementation of IGBTs in robotics was in General Motors's Saturn plant in Tennessee.

In the highest-voltage category are applications in locomotive drives and in electrical-power distribution and transmission, including conversion between alternating-current and direct-current electricity. Voltage ratings can exceed 5,000 volts, and the devices must be capable of handling 1,000 amperes. A kind of semiconductor device known as a thyristor is commonly used to handle such high voltages and currents. Yet IGBTs have just recently begun capturing the lower end of this category, thanks to the introduction, by several Japanese companies, of devices capable of operating at 3,500 volts and 1,000 amperes. Such high-power IGBTs are now controlling the motors of Japan's bullet trains, among other things.

Best of Both Worlds

IGBTs are a wonderful example of a whole that is much more than the sum of its parts. Each IGBT consists of two transistors: a MOSFET and another kind, known as bipolar. Bipolar transistors are the simplest, most rugged type of transistor, having evolved directly out of the pioneering work at Bell Telephone Laboratories in the late 1940s. They can be designed to accommodate high power levels and can be switched on and off at extremely high speeds. Unfortunately, they require a fairly substantial flow of electric current in order to control a larger current. (A more succinct way of saying this is that their power gain is modest.) MOSFETs, on the other hand, are unable to handle high power levels but have fabulous gain. Through clever design, the IGBT combines the best features of these two different devices.

The way in which the IGBT accomplishes this trick is rather impressive—and the result of years of intensive research at General Electric's research laboratories. This achievement cannot be

Copyright 1997 Scientific American, Inc.

understood without some background on transistor operation and on the way in which the bipolar and MOSFET varieties work.

Transistors can be designed to operate as switches, blocking or permitting the flow of electric current, or as amplifiers, making a minute current much greater. In power electronics, where engineers are concerned mainly with switching, transistors are distinguished by the amount of power they can control.

Electricity is analogous to liquid flow in a pipe. Just as hydraulic power is the product of pressure and volumetric flow rate, electrical power is the product of voltage and current. Thus, the amount of power that a transistor can control is decided by its maximum operating voltage and current handling capability. In its "on" state, the transistor allows current to flow through itself and be delivered to the load, which might be a heater, a motor winding or some other system. In the "off" state, the transistor stops current flow by supporting a high voltage without letting current through.

Transistors typically have three electrical leads, which are also called terminals. The relatively large "working" current that flows to the load passes through the transistor between the terminals connected to the parts of the transistor referred to as the emitter and the collector; in the MOSFET, these parts are the source and the drain. The smaller "control" current, which turns the working current on and off, flows between the third part (the base, or gate in the MOSFET) and the emitter (or source).

HEARTSTREAM

ADVANCED DEFIBRILLATOR based on IGBTs delivers a heart-starting jolt through paddles applied to the chest.

The emitter, base and collector are separate sections of the transistor. Each is made of a material, typically silicon, that has been impregnated, or "doped," with impurities to give it certain desired electrical properties. If the doping gives the material an excess of mobile electrons, the material is called *n*-type. Conversely, if the material has been doped to have deficiencies of electrons (which are known as holes), it is designated *p*-type. A bipolar transistor is created by sandwiching three layers of these semiconductor types, in the order *n-p-n* or, alternatively, *p-n-p*. In other words, the emitter can be *n*-type, in which case the base is *p*-type, and the collector is *n*-type.

In the *n-p-n* bipolar transistor, the working current flows from the emitter across the base to the collector—but only when the control current is flowing. When it flows, the control current adds holes to the base, thereby attracting electrons from the emitter. When the control current stops, holes are no longer added to the base, and the working current stops flowing. The operation of a *p-n-p* transistor is essentially identical to that of the more common *n-p-n* with one important difference: in the *p-n-p,* the roles of electrons and holes are reversed with respect to the *n-p-n* [*see illustration on preceding two pages*].

Block That Current

The ability of a transistor to prevent current from flowing, even when a high voltage is applied across its emitter and collector terminals, is one of the most basic requirements in power electronics. This characteristic is achieved by varying the size and dopant concentrations of the transistor's regions, particularly the collector.

To understand how this feature is achieved, consider how a transistor blocks the flow of current. Current is blocked near the interface, or junction, where *p*-type material and *n*-type material meet—for example, at the junction between the base and the collector. Suppose the relatively positive voltage is connected to the *n*-type material, and the relatively negative voltage is connected to the *p*-type material. The junction is said to be reverse biased, and it can block current from flowing. Specifically, the reverse biasing creates on either side of the junction so-called depletion regions, where a lack of electrons (holes) makes it impossible for current to flow.

Working against this depletion region, which is blocking the flow of current, is an electrical field in the collector. In effect, this field promotes the flow of current, because the electrons move through the collector under the influence of the field. As the voltage is increased, the field becomes more intense, until finally the resistance offered by the depletion region is overcome, and current flows across the junction. Thus, it is important to minimize this field, which is done by making the collector relatively thick and doping it very lightly. With these techniques, junctions have been made that can withstand the application of thousands of volts.

In contrast to the thick, lightly doped collector, the base in a bipolar transistor is thin and heavily doped. These features promote the diffusion of electrons through the base, which is needed to ensure good current-carrying capacity when the transistor is in the on state. The ability to conduct large currents is also necessary to ensure that the device has a sufficiently high power gain, which is defined as the ratio of the power being controlled, in the collector-emitter circuit, to the input power, in the base-emitter circuit. The power being controlled (in the collector-emitter circuit) is the product of the current carried by the device in its on state and the voltage at the collector terminal when the device is in its off state.

In a bipolar transistor, a base current can control a current flow in the collector that is about 10 times greater. Typically, the voltage on the collector is not quite 100 times that at the base, so bipolar transistors operate at a power gain of less than 1,000. One implication of this modest gain is that at the kilowatt levels of most power-electronics applications, the control circuit must be capable of handling several watts. This level of power in turn demands a relatively complex and robust control circuit. Moreover, for completely safe operation, bipolar transistors must be used with other protective components.

The success of the superior MOSFET and IGBT devices is increasingly pushing bipolar transistors into what might be called niche markets. Still, thyristors, which actually comprise a pair of bipolar transistors, dominate the highest-voltage applications. Thyristors are available that can support 6,000 volts in the off state and carry 1,000 amperes in the on state. In other words, a single semiconductor device—a wafer 10 centimeters in diameter—is capable of controlling six megawatts of power! Unfortunately, the current gain of these de-

Copyright 1997 Scientific American, Inc.

vices is less than five, requiring an enormous control current supplied by complex, bulky and heavy control circuits that make them inconvenient to use, for example, in vehicles. In addition, the maximum switching speed of these thyristors is so slow that the system operating frequency is within the range of human hearing, resulting in noise pollution that arises from vibrations in the motor windings.

How MOSFETs Work

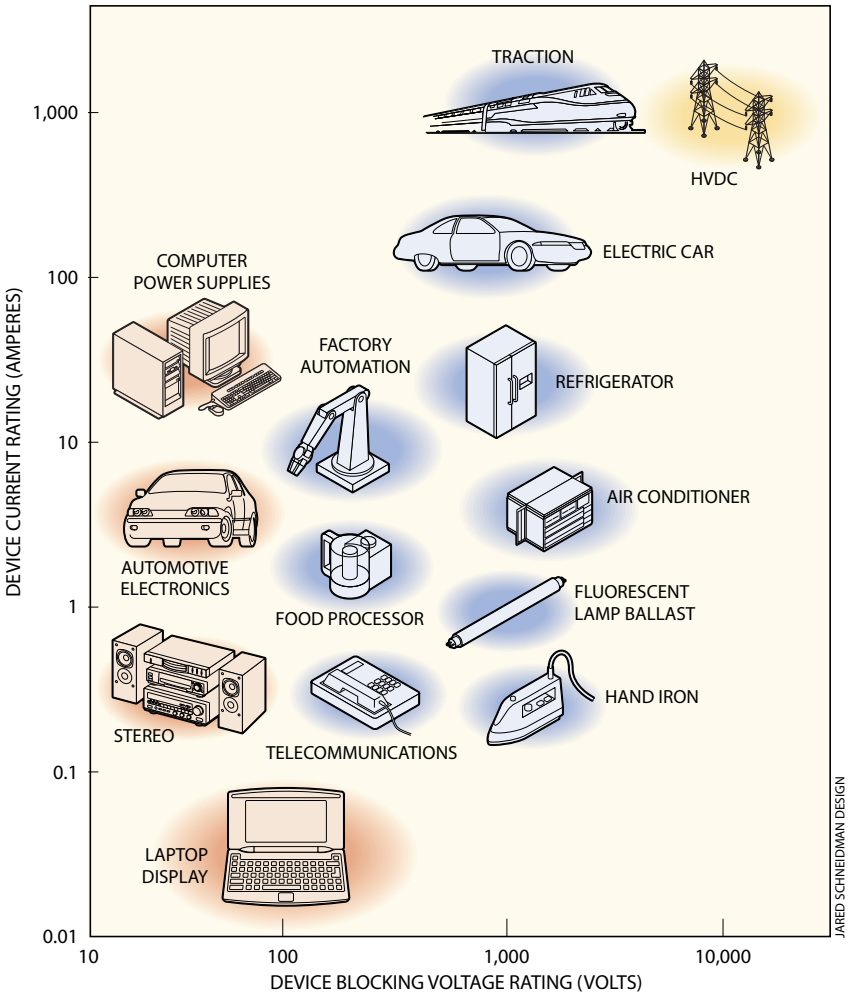
In comparison with the bipolar transistor, the other component of the IGBT, the MOSFET, operates on a rather different principle. An *n-p-n* MOSFET (more correctly termed an *n*-channel MOSFET) has two *n*-type regions—the source and the drain—which are analogous to the bipolar transistor's emitter and collector. In between the source and drain is a *p*-type region, called the substrate [*see illustration on page 36*].

On top of the substrate, which is made of a silicon semiconductor material, is a nonconductive layer of silicon dioxide; on top of this oxide layer is a metal "gate." (Hence the first three letters of the acronym stand for metal oxide semiconductor.) Normally, no charges flow from the source through the substrate, immediately below the oxide layer, to the drain. When a positive voltage is applied to the metal gate, however, an electrical field is set up that penetrates through the oxide layer and into the substrate. (Hence the second three letters of the acronym: field-effect transistor.) This field repels the positively charged holes (electron deficiencies) in the substrate, forcing them from the gate. At the same time, it attracts the electrons toward the substrate surface, just below the oxide layer. These mobile electrons then allow current to flow through the substrate, just below the oxide, between the drain and the source.

The most important aspect of the MOSFET's operation is the fact that it is turned on and off with voltage, not current. Current flow through the gate is limited to short, milliampere pulses that occur only when the transistor is turned on or off. (These pulses occur because the semiconductor substrate and metal gate, separated by the oxide layer, form a capacitor that causes transient currents when the capacitor charges and discharges.)

MOSFETs are an offshoot of the complementary metal oxide semiconductor (CMOS) technology developed in the early 1970s for microelectronics. In fact, CMOS technology now forms the basic building block for all commercially available silicon integrated circuits. Although makers of power transistors relied on bipolar technology at that time, engineers realized that they could increase the power gain of transistors by exploiting the MOS-gate structure.

This realization led to the development of power MOSFETs in the 1970s by International Rectifier Corporation in El Segundo, Calif. Besides having higher power gain, the devices switched faster and did not require the cumbersome protection circuits used with bipolar transistors. Though ideal in many ways, power MOSFETs do have one major drawback: their current-handling capability degrades rapidly when they are designed to operate at more than 100 volts. Above this voltage level, the electrical resistance inside the device begins to soar, severely limiting the current that can be coaxed out of the drain.



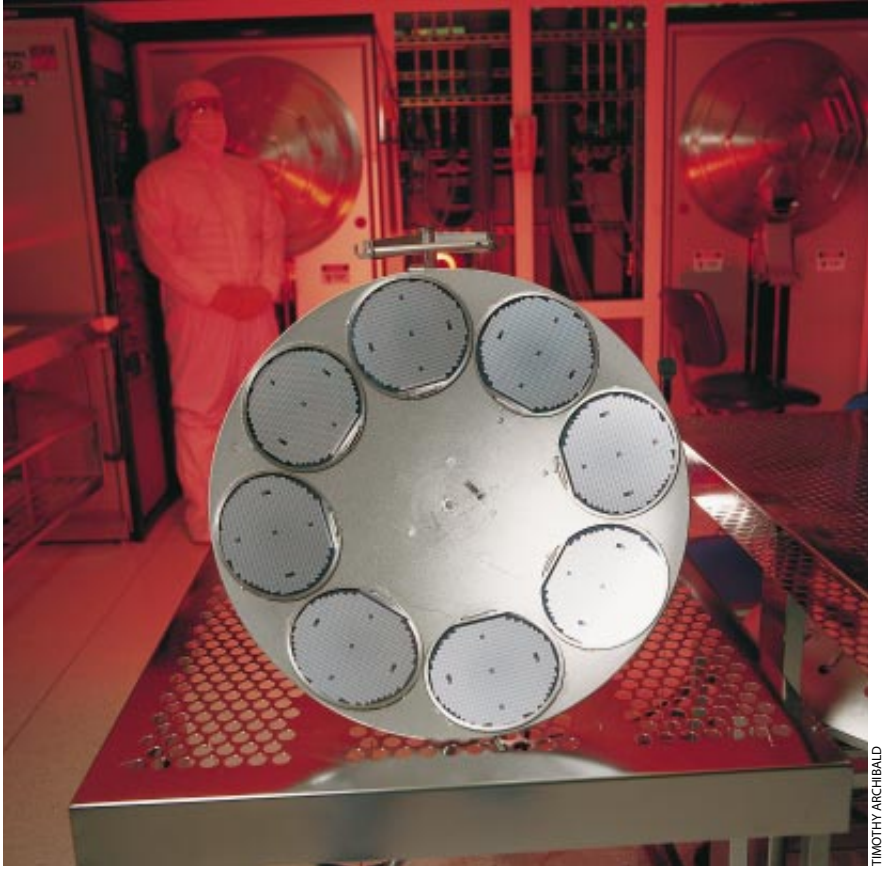
VOLTAGE AND CURRENT RATINGS needed for different power transistor uses vary considerably. The lowest-voltage applications (*highlighted in pink*) are still served mostly by MOSFETs. On the other end, with the greatest voltage-blocking and current-handling capabilities, are the thyristors used in high-voltage, direct-current (HVDC) electrical transmission systems. The large set of intermediate applications (*blue*) is increasingly dominated by IGBTs.

MOSFET + Bipolar = IGBT

In the late 1970s, while I was working for General Electric's research laboratory in Schenectady, N.Y., I had the idea of integrating MOSFET and bipolar technologies into one device. With the MOSFET controlling the bipolar, I reasoned, the integrated device could be switched by tiny voltages and yet allow hundreds of amperes to flow through it. Ultimately, this realization led to the IGBT, but the path was not direct.

The first MOS-bipolar power device I built at GE, in 1978, was not an IGBT but rather a MOS-gated thyristor. This device, which became a commercial product that is still available today, can

Copyright 1997 Scientific American, Inc.



TIMOTHY ARCHIBALD

CENTRIFUGAL DOME holds eight silicon wafers, each of which will yield roughly 320 IGBTs. In a vacuum chamber, precious metals are sputtered onto the back of the wafers to produce the conductive components of each device. The assembly was photographed at Motorola's MOS 4 power-transistor fabrication facility in Phoenix, Ariz.

deliver a pulse of current from a capacitor to, for example, a gas discharge tube of the kind used in photolithography tools. Engineers continue to work on MOS-gated thyristors in hopes of producing a device capable of replacing a kind of thyristor, known as a gate-turn-off thyristor, commonly used in the highest power applications.

In an IGBT, there is just one useful bipolar transistor (as opposed to the pair that comprise a thyristor), and it is a *p-n-p* type. This *p-n-p* transistor is a rather unusual one in several respects. For one, the typical, commonly available bipolar power transistor is an *n-p-n,* not a *p-n-p*. Moreover, the typical power transistor has a narrow base region and a thick, lightly doped collector. As mentioned before, the thin base enables large currents to flow through it in the on state, whereas the thick, lightly doped collector blocks current in the off state.

In the *p-n-p* transistor in an IGBT, on the other hand, the characteristics of base and collector are reversed: the base is thick and lightly doped; the collector is relatively thin and very highly doped. How is this reversal possible? Think back to the key requirements of a power semiconductor device. One is that in the off state, the device must be able to support a high voltage across its output terminals, the emitter and the collector. In a conventional power transistor, this requirement is satisfied by making the collector thick and lightly doped. But a thick collector and a thin base were found to be impractical in the IGBT for reasons having to do with chip fabrication and performance limitations.

Fortunately, it is a fact that the voltage in a power transistor can be blocked by making *either* the base *or* the collector thick and lightly doped. The reason the collector is invariably made thick in a conventional power transistor is that high current gain demands a thin, highly doped base.

But what if we do not care about current gain in the bipolar transistor? I realized that this is precisely the case with the IGBT, because it is the MOSFET, with its huge current gain, that provides the control current to the bipolar transistor. In other words, the two parts of an IGBT are integrated together in such a way that the channel current flowing in the substrate of the MOSFET is also the current that is applied to the base of the bipolar power transistor. Thus, so much current is being provided to the base of the bipolar transistor that low amplification (typically by a factor between one and two) suffices.

As mentioned previously, there are small current transients when the IGBT's MOSFET is switched on and off, resulting in short pulses of current on the order of milliamperes. This MOSFET is controlled by voltages on the order of 10 volts, and the IGBT is capable of controlling 1,500 volts and 100 amperes. Using these values, it is possible to calculate that the power gain of an IGBT exceeds 10 million.

Such high gain not only enables the IGBT to be controlled by relatively delicate integrated circuits (ICs), it also permits the inclusion of protection circuits in the control IC to prevent destructive failure. Such failures are a distinct possibility when the device is misused—for example, when it is operated beyond its specified temperature, current capacity or voltage level.

Another attribute of the IGBT is its significantly higher operating current density in the on state when compared with its two components, the bipolar transistor and the MOSFET. Recall that the current flowing in the channel of the MOSFET is used as the input, or control, current for the bipolar. Because of the way the two transistors are integrated together, the output current of the IGBT consists not just of the bipolar's emitter-collector current, as might be expected, but of the sum of that current and the channel current in the MOSFET. These two currents are roughly equal (the gain of the bipolar is only about one or two), so the output current of the IGBT is approximately twice that of either of its components.

Another important feature that enhances the efficiency of the IGBT is its unusually low electrical resistance, in the on state, between its emitter and collector. This property comes from the large concentration of electrons and holes that are injected into the bipolar's wide, lightly doped base region from the adjacent emitter and collector during current flow. This flooding of charge carriers increases the base's conductivity 1,000 times. Therefore, the power losses inside the device are exceptionally low in comparison with ordinary MOSFETs or even bipolars. For any particular application, this feature translates into a proportionate reduction in chip area, which in turn leads to a substantial reduction in the cost for manufacturing the device.

Copyright 1997 Scientific American, Inc.

The main difficulty with introducing an IGBT commercially was the existence in the device of a so-called parasitic thyristor. This thyristor arises from the presence of four adjacent semiconductor layers, alternately *p*-type and *n*-type. These layers form two de facto bipolar transistors (one *n-p-n* and one *p-n-p*) with a common collector junction that enables them to feed back between each other. The condition leads to destructive failure of the device. The problem was solved through a combination of structural innovations, including the addition of another highly doped *p*-type region under the MOSFET's *n*-type source region.

IGBTs on the Move

The rapid adoption of IGBTs throughout most of the various categories of power electronics shows no sign of slowing down. One category with plenty of room for expansion is transportation. In addition to the benefits of smaller size and weight for these transportation systems, IGBT-based power electronics are capable of operating at a higher frequency. Several Japanese companies, including Fuji Electric, Mitsubishi Electric, Hitachi and Toshiba, have shown that this higher-frequency operation makes for a smoother ride and a quieter passenger cabin. Plans to implement IGBT-based electric streetcars and locomotives are under way in Europe; the corresponding IGBT developments are going on at ABB Corporation and Siemens.

Electric and hybrid-electric automobiles are the subject of intense development lately, as a consequence of concerns about the environmental pollution resulting from gasoline-powered internal-combustion engines. Most electric and hybrid cars condition and convert the direct current of their batteries to alternating current for the motor with IGBT-based systems, called inverters. IGBTs are also used to convert alternating current to direct current to recharge the batteries; this conversion must be highly and precisely regulated to avoid damaging the battery electrodes.

More efficient control of electrical power, through the use of IGBTs, will enhance living standards.

Cars and trains are not the only electric vehicles that will benefit from the precision and power of IGBTs. As part of an effort to reduce urban pollution, the Shanghai Energy Commission in China will produce 150,000 electric mopeds in 1998 while restricting the sale of gasoline-powered models. The widespread introduction of these electric vehicles will demand an effective means for charging the batteries either rapidly at roadside stations or overnight at home.

Perhaps the most gratifying use of the IGBT will be in the saving of thousands of lives every day around the world. Every year more than 350,000 people die of sudden cardiac arrest in the U.S. alone because the only effective treatment, an external defibrillator, is not immediately accessible. The size and weight of these systems, which deliver an electric shock to restart the patient's heart, have been significant stumbling blocks to their wider deployment. Now, however, a Seattle-based medical concern called Heartstream is marketing a compact, lightweight defibrillator based on IGBTs. Heartstream's system, which was approved by the U.S. Food and Drug Administration in 1996, is starting to replace the larger units now carried in emergency vehicles and on commercial airliners. Last July, for example, American Airlines announced that it had equipped 247 of its aircraft with the Heartstream units. The American Heart Association estimates that 100,000 lives could be saved in the U.S. if defibrillators were more widely available.

IGBTs already have many other medical uses, albeit less dramatic ones. They are an essential component of the uninterruptible power supplies used in hospitals to ensure fail-safe operation of medical equipment during brief power outages. In addition, IGBTs drive the motors in computed tomographic (CT) scanners for the precise movement of the x-ray unit to produce sectional images of a body.

Applications such as these are good examples of technology doing what it is supposed to do: serve humanity. At the same time, the uses all point to a wonderful, uncommon occurrence—the coming together of two technologies to accomplish what neither could do by itself. The invention of the transistor and its miniaturization have led to complex integrated circuits that can be used to process information in digital form. The transistor triggered the first electronic revolution, which has brought us into the information age. Yet the efficient and effective control and utilization of electrical power are essential for enhancement of our living standards by giving us better control over our appliances, our living environments and our vehicles. The invention and rapid commercialization of the IGBT have played a major role in making that control possible. The second electronic revolution is upon us, and before it is over we will all benefit from it. SA

The Author

B. JAYANT BALIGA is director of North Carolina State University's Power Semiconductor Research Center, which he founded in 1991. From 1979 to 1988 he was manager of power-device development at the General Electric Corporate Research and Development Center in Schenectady, N.Y. He has been a professor of electrical engineering at North Carolina State since 1988. His current research interest lies in developing silicon and silicon-carbide-based power semiconductor devices.

Further Reading

EVOLUTION OF MOS-BIPOLAR POWER SEMICONDUCTOR TECHNOLOGY. B. Jayant Baliga in *Proceedings of the IEEE,* Vol. 76, No. 4, pages 409–418; April 1988.

POWER ICS IN THE SADDLE. B. Jayant Baliga in *IEEE Spectrum,* Vol. 32, No. 7, pages 34–49; July 1995.

POWER SEMICONDUCTOR DEVICES. B. Jayant Baliga. PWS Publishing, Boston, 1996.

TRENDS IN POWER SEMICONDUCTOR DEVICES. B. Jayant Baliga in a special issue of *IEEE Transactions on Electron Devices,* Vol. 43, No. 10, pages 1717–1731; October 1996.

Copyright 1997 Scientific American, Inc.

WHERE TUBES RULE

In a solid-state world, vacuum tubes are still secure within some very interesting strongholds

by Michael J. Riezenman

Passion" and "cult" aren't words one normally associates with electrical engineering, but they do come into play the minute tubes are mentioned.

Vacuum tubes, that is to say. Audio tubes, to be precise.

For the uninitiated, which you almost certainly are if you're younger than 40 years of age, vacuum tubes were the active electronic devices used by primitive peoples before transistors and integrated circuits were invented. In fact, in the early part of this century, the very word "electronics" referred to a branch of physics concerned with the behavior of electrons in a vacuum.

Devised in 1904 by John Ambrose Fleming, the first tubes (or valves, as the British call them) were simple diodes, which permitted electric current to flow in only one direction. Electronics really took off around 1912, when Edwin Howard Armstrong figured out how to build useful amplifier and oscillator circuits with the audion tube, invented six years earlier by Lee De Forest. By inserting an electrode known as a grid between the diode's other two electrodes, known as the cathode and anode, De Forest created a controllable device in which small changes in the voltage on the grid resulted in larger changes in the current flowing between the cathode and the anode. Such a three-electrode tube is called a triode.

RADIO-FREQUENCY TRANSMITTING TUBES such as the 845 (*left*) and the 811A (*above, right*) were used as power-amplifying devices in 1940s-era amateur radio and broadcast transmitters. In the 811A, as in many transmitting types, a high, direct-current voltage is applied through a plate cap on top of the tube and modulated inside the tube by an input signal.

Although the evidence today seems to suggest that DeForest had only a slight appreciation of what he had wrought, after much experimentation, Armstrong did. In a seminal moment in electronics history, he coupled the tube's output circuit back to its input to boost its feeble gain, thereby inventing the positive feedback circuit.

Over time, thousands of different tubes were developed, from subminiature devices the size of a cigarette filter to the hefty units still used in high-power radio transmitters, radar and industrial heating equipment. In addition to triodes, engineers came up with tetrodes, pentodes and other tubes with multiple-grid electrodes.

Small receiving tubes, of the kind found in tabletop radios by the millions between about 1920 and 1960, have now been almost completely displaced by transistors, which seem to last forever. They require neither high voltages nor warm-up time, lend themselves to real miniaturization and use far less power.

Pleasure and Passion

So pervasive have transistors become that few people today even think about tubes in the context of home audio equipment. There exists, however, a small but passionate minority that believes that the best transistor-based amplifiers cannot re-create a piece of music as pleasingly as can a properly designed amplifier built around vacuum triodes. "Pleasing," of course, is a subjective word, and that is where the passion comes in.

As explained by Kevin M. Hayes, founder and president of Valve Amplification Company in Durham, N.C., a manufacturer of tube-based audio amplifiers, the case for tubes begins with the realization that industry-standard laboratory measurements of amplifier performance do not adequately answer fundamentally subjective questions such as "Is this amplifier better than that one?" The problem, he says, is that the workings of the ear and brain are not understood well enough to identify the necessary and sufficient set of measurements for answering the question.

Back in the 1930s and 1940s, total harmonic distortion became a widely accepted parameter for describing amplifier imperfections. All amplifiers create spurious, ostensibly unwanted signals at frequencies that are some whole-number multiple of the signal being amplified. Thus, a second-order harmonic distortion consists of stray signals

Copyright 1997 Scientific American, Inc.

at exactly twice the frequency of the amplified signal. Because all amplifiers of that era were based on tubes with similar kinds of nonlinearities, they all tended to generate harmonic distortion of the same kind, and a single number representing the total harmonic distortion was a valid tool for comparing them. It correlated well with the subjective listening experience.

Those tube amplifiers generated mainly second-order harmonics plus small amounts of other low-order even harmonics (fourth, sixth and so on). Second-order harmonic distortion, for that matter, is difficult for a human being to detect. Moreover, what can be heard tends to sound pleasant.

Transistor amplifiers, in contrast, generate higher-order harmonics (ninth, tenth, eleventh and so on), which are much easier to hear. Worse, the odd-order ones sound bad. So it is possible to have a transistor amplifier whose total harmonic distortion—as measured by laboratory instruments—is significantly lower than that of a comparable tube amplifier but that nonetheless sounds worse. To make a long story short, total harmonic distortion is not a particularly good way to compare amplifiers based on fundamentally different technology, and it is not clear what is—other than listening to them, of course.

The debate can get quite heated—and not a little confusing—because the performance of an amplifier depends as much on the details of its circuit design as on its principal active devices (tubes or transistors). For example, using the feedback configuration in an amplifier circuit can reduce total distortion levels, but at a price: an increased percentage of those easily perceived, higher-order harmonics. According to Hayes, transistor amplifiers need more feedback than amplifiers based on vacuum triodes, which he believes to be the most optimal audio-amplifying devices. (Hayes's favorite triode is the 300B, developed at Western Electric in 1935.)

Tube Strongholds

Then there are the cultists who not only prefer tube-based audio equipment but insist that single-ended amplifiers are superior to push-pull units. In the latter, pairs of output tubes are arranged in a circuit that tends to cancel even-order distortion. Single-ended outputs, lacking that cancellation, can have as much as 15 percent total harmonic distortion, mostly second order. Though readily detectable, the effect is not unpleasant, tending to add richness and fullness to the reproduced sound. According to Hayes, it was used deliberately by manufacturers to improve the tinny sound quality of 1940s radios.

Fraught as it is with human and technical interest, the controversial audio market is a relatively tiny part of the tube business, which is far larger than most people imagine. Tubes are still playing a major role in high-frequency, high-power applications. In general, at almost any frequency there is a power level above which it makes more sense to use a tube rather than an array of transistors as the final power amplifier.

Microwave ovens are a good case in point. They need to put out a few hundred watts at two or three gigahertz, a requirement easily satisfied by a kind of tube known as a magnetron, which costs about $18 to $25 in quantity. These microwave oven magnetrons descended from those used since the earliest days of radar, during World War II. (Remarking on the importance of radar during the war, Winston Churchill once described an early magnetron as the most valuable cargo ever to cross the Atlantic Ocean.)

Radio transmitters are another point of interest in tube country. In building power amplifiers for AM radio transmitters, where the goal is to generate 25 or 50 kilowatts at one or two megahertz, the tendency today is to go solid-state. For ultrahigh-frequency transmitters, which operate above 300 megahertz, tubes still reign supreme.

State-of-the-art communications satellites, too, are typically tube-equipped. Intelsat—the international consortium that operates about half the world's commercial communications satellites—employs both solid-state and tube-based power amplifiers in its Series VIII satellites, which are just now going into service. Their Ku-band transmitters, which work at around 12 gigahertz and generate fairly narrow "spot" beams of ground coverage, use amplifiers based on so-called traveling wave tubes. The lower-frequency C-band transmitters operate at about five gigahertz and use both technologies, depending on how much power they are intended to deliver. Below about 40 watts, they use arrays of gallium arsenide field-effect transistors. Above that level, it's traveling wave tubes.

Although predicting the future is a notoriously tricky business, especially when it comes to electrotechnology, it seems safe to say that tubes will be with us for a long time. Undoubtedly, the frequency and power levels that can be handled by solid-state amplifiers will keep climbing. But the infinite space above them will almost certainly remain tube territory. SA

PHOTOGRAPHS BY CORDERO STUDIOS

MICHAEL J. RIEZENMAN is senior engineering editor of *IEEE Spectrum* magazine. During the 1960s, as an electrical engineer with ITT Defense Communications, he designed control circuitry for use in communications satellites.

Copyright 1997 Scientific American, Inc.

As the transistor has grown smaller and cheaper, engineers have scoffed at theoretical barriers to its progress—so far

THE FUTURE OF THE TRANSISTOR

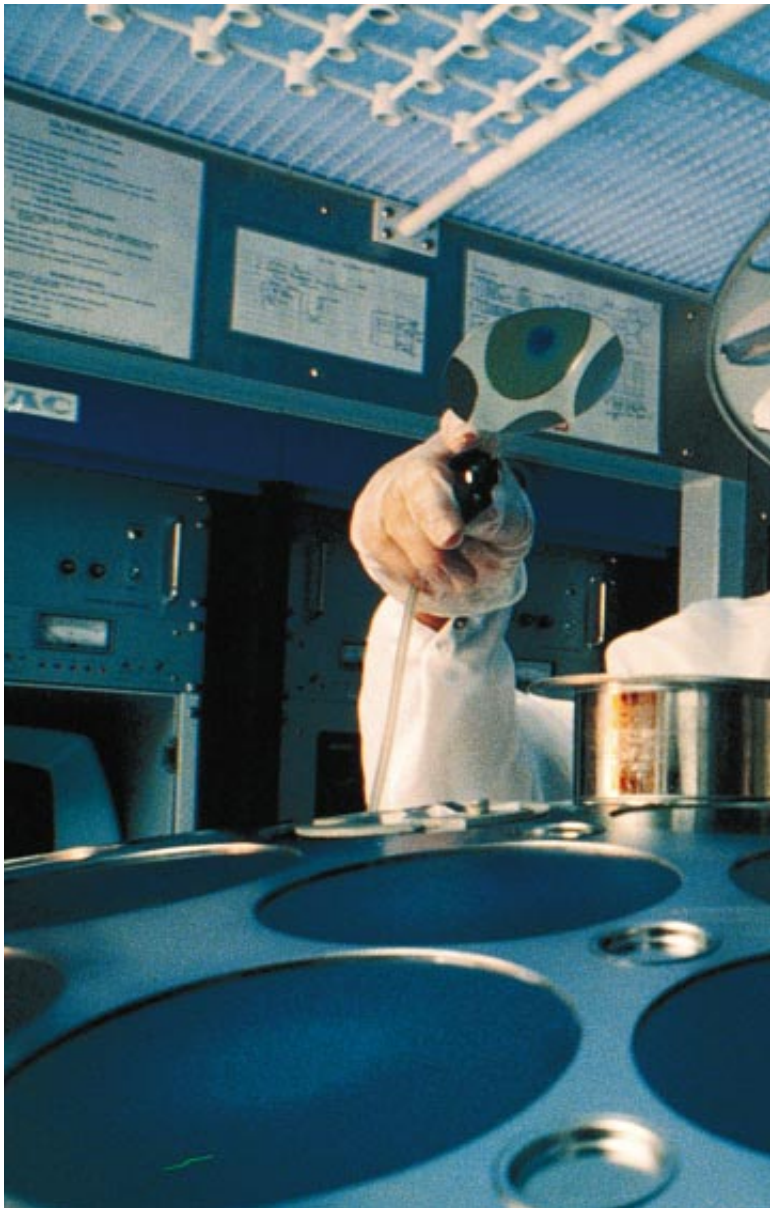
by Robert W. Keyes

I am writing this article on a computer that contains some 10 million transistors, an astounding number of manufactured items for one person to own. Yet they cost less than the hard disk, the keyboard, the display and the cabinet. Ten million staples, in contrast, would cost about as much as the entire computer. Transistors have become this cheap because during the past 40 years engineers have learned to etch ever more of them on a single wafer of silicon. The cost of a given manufacturing step can thus be spread over a growing number of units.

How much longer can this trend continue? Scholars and industry experts have declared many times in the past that some physical limit exists beyond which miniaturization could not go. An equal number of times they have been confounded by the facts. No such limit can be discerned in the quantity of transistors that can be fabricated on silicon, which has proceeded through eight orders of magnitude in the 50 years since the transistor was invented [*see box on pages 50 and 51*].

I do not have a definitive answer to the question of limits. I do, however, have some thoughts on how the future of solid-state electronics will develop and what science is needed to support continuing progress.

Several kinds of physical limitations might emerge as the size of the transistor continues to shrink. The task of connecting minute elements to one another might, for example, become impossible. Declining circuit size also means that researchers must cope with ever stronger electrical fields, which can affect the movement of electrons in many ways. In the not too distant future the transistor may span only hundreds of angstroms. At that point, the presence or absence of single atoms, as well as their behavior, will become significant. Diminishing size leads to increasing density of transistors on a chip, which raises the amount of waste heat thrown off. As the size of circuit elements drops below the wavelength of usable forms of radiation, exist-



MINIATURIZATION has made transistors cheaper than staples by spreading manufacturing costs over millions of devices on each of the hundreds of chips on a wafer. This worker holds a nearly completed wafer. Its components will be connected by the condensation of metal in a vacuum chamber (*foreground*).

Copyright 1997 Scientific American, Inc.

ing manufacturing methods may reach their limits.

To see how such problems might arise and how they can be addressed, it is useful to review the operation of the field-effect transistor, the workhorse of modern data processing. Digital computers operate by manipulating statements made in a binary code, which consists of ones and zeroes. A field-effect transistor is operated so that, like a relay, it is switched only "on" or "off." The device therefore represents exactly one binary unit of information: a bit. In a large system, input signals control transistors that switch signal voltages onto output wires. The wires carry the signals to other switches that produce outputs, which are again sent on to another stage. The connections within the computer determine its function. They control the way that the inputs are transformed to become outputs, such as a word in a document or an entry in a spreadsheet.

From Source to Drain

The field-effect transistor contains a channel that interacts with three electrodes: a source, which supplies electrons to the channel; a drain, which receives them at the other side; and a gate, which influences the conductivity of the channel [*see illustration on next page*]. Each part contains different impurity atoms, or dopants, which modify the electrical properties of the silicon.

The gate switches the transistor on when a positive voltage applied to it attracts electrons to the interface between the semiconductor and the gate insulator. These electrons then establish a connection between the source and drain electrodes that allows current to be passed between them. At this point, the transistor is "on." The connection persists for as long as the positive charge remains on the gate. An incoming signal is applied to the gate and thus determines whether the connection between source and drain is established. If a connection results, the output is connected to the ground potential, one of the standard digital voltages. If no connection results, the output is connected through the resistor to the positive power supply, the other standard digital voltage.

Circuits of transistors must be oblivious to the operations of neighboring arrays. Existing concepts of insulation,

IBM CORPORATION

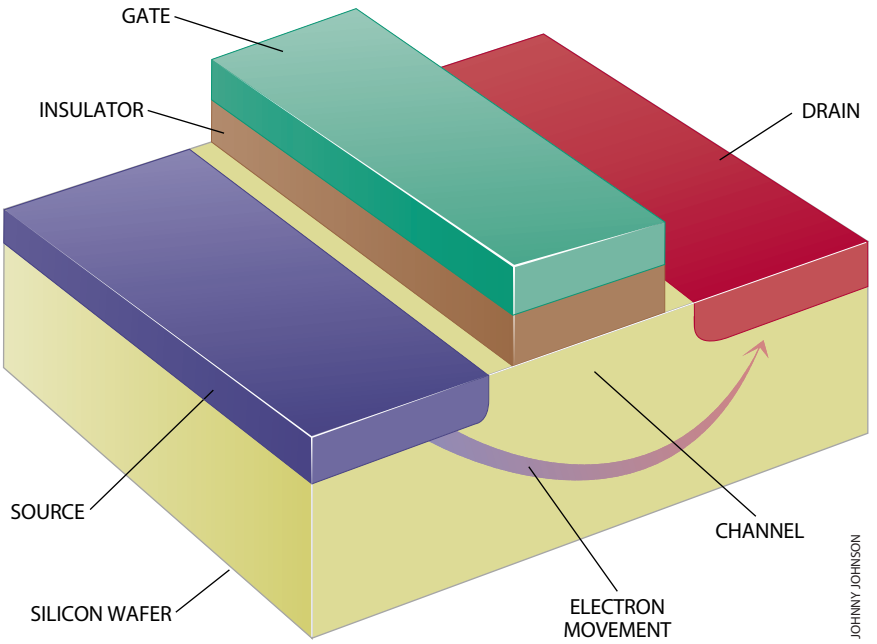
Copyright 1997 Scientific American, Inc.

impedance and other basic electrical properties of semiconductors and their connections should work well enough, for designers' purposes, in the next generation of devices. It is only when conducting areas approach to within about 100 angstroms of one another that quantum effects, such as electron tunneling, threaten to create problems. In laboratory settings, researchers are already at the brink of this limit, at about 30 angstroms; in commercial devices, perhaps a decade remains before that limit is reached.

Another challenge is the strengthening of the electrical field that inevitably accompanies miniaturization. This tendency constrains the design of semiconductor devices by setting up a basic conflict. Fields must continually get stronger as electron pathways shrink, yet voltages must remain above the minimum needed to overwhelm the thermal energy of electrons. In silicon at normal operating temperatures, the thermal voltage is 0.026 electron volt. Therefore, whenever a semiconductor is switched so as to prevent the passage of electrons, its electrical barrier must be changed by a factor several times as large. One can minimize the thermal problem by chilling the chip (which becomes an expensive proposition).

It has only recently been taken for granted that anyone can search for references to anything from kiwifruit to quantum physics.

Even cooling cannot end the problem of the electrical field. Signals must still have the minimum voltage that is characteristic of a semiconductor junction. In silicon this electrical barrier ranges between half a volt and a volt, depending on the degree of doping. That small voltage, applied over a very short distance, suffices to create an immensely strong electrical field. As electrons move through such a field, they may gain so much energy that they stimulate the creation of electron-hole pairs, which are themselves accelerated. The resulting chain reaction can trigger an avalanche of rising current, thereby disrupting the circuit. Today's chips push the limits in the quest for high speed, and electrical fields are usually close to those that can cause such avalanches.



FIELD-EFFECT TRANSISTOR, the workhorse of data processing, is built as a sandwich of variously doped silicon layers. It contains a channel, a source, a drain and an insulated gate. When a positive voltage is applied to the gate, electrons move near the insulation, establishing a connection underneath it that allows current to pass from source to drain, switching the transistor on.

Tricks and Trade-offs

Workers resort to a variety of tricks to mitigate the effects of strong electrical fields. They have designed field-effect transistors, for example, in which the field can be moved to a place where it does not disrupt other electronic functions. This stratagem is just one of many, all of which entail trade-offs with other desired characteristics, such as simplicity of design, ease of manufacture, reliability and long working life.

Miniaturization also increases the heat given off by each square centimeter of silicon. The reason is purely geometric: electrical pathways, and their associated energy losses, shrink in one dimension, whereas chip area shrinks in two. That relation means that as circuits get smaller, unit heat generation falls, albeit more slowly than does the number of units per square centimeter.

Devices already pour out as much as 30 watts per square centimeter, a radiance that one would expect of a material heated to about 1,200 degrees Celsius (this radiance value is about 10 times that of a range-top cooking surface in the home). Of course, the chips cannot be allowed to reach such temperatures, and so cooling systems remove heat as fast as it is produced. A variety of cooling technologies have been devised, including some rather intense ones. But the cost of using them in transistor circuits increases rapidly when the density of heat increases.

The exigencies of manufacturing impose constraints on the performance of electronic devices that might not be apparent from a purely theoretical discussion. Low-cost manufacturing results in small differences among the devices that are made on each wafer, as well as among those that are fabricated on different wafers. This variability cannot be banished—it is inherent in the way solid-state devices are made.

A semiconducting material, such as silicon, is made into a transistor in an integrated process involving many steps. Templates, called masks, are applied to the silicon in order to expose desired areas. Next, various operations involving chemical diffusion, radiation, doping, sputtering or the deposition of metal act on these areas, sometimes by constructing device features, other times by erect-

Copyright 1997 Scientific American, Inc.

ing scaffolding to be used in succeeding steps and then torn down. Meanwhile other devices—resistors, capacitors and conductors—are being built to connect the transistors.

Variations intrude at every step. For example, perfect focusing of the source of radiation over a large wafer is hard to achieve. The temperature of the wafer may vary slightly from one place to another during processing steps, causing a difference in the rate of chemical reactions. The mixing of gases in a reaction chamber may not be perfect. For many reasons, the properties of devices on a given wafer and between those on different wafers are not identical. Indeed, some devices on a wafer may be no good at all; the proportion of such irremediable errors places a practical limit on the size of an integrated circuit.

A certain amount of fuzziness is inherent in optical exposures. The light used in photolithography is diffracted as it passes through the holes in the template. Such diffraction can be minimized by resorting to shorter wavelengths.

When photolithographic fabrication was invented in the early 1970s, white light was used. Workers later switched to monochromatic laser light, moving up the spectrum until, in the mid-1980s, they reached the ultraviolet wavelengths. Now the most advanced commercial chips are etched by deep ultraviolet light, a difficult operation because it is hard to devise lasers with output in that range. The next generation of devices may require x-rays. Indeed, each generation of circuitry requires manufacturing equipment of unprecedented expense.

Other problems also contribute to the cost of making a chip. The mechanical controls that position wafers must become more precise. The "clean rooms" and chambers must become ever cleaner to ward off the ever smaller motes that can destroy a circuit. Quality-control procedures must become even more elaborate as the number of possible defects on a chip increases.

Device "Sandwich"

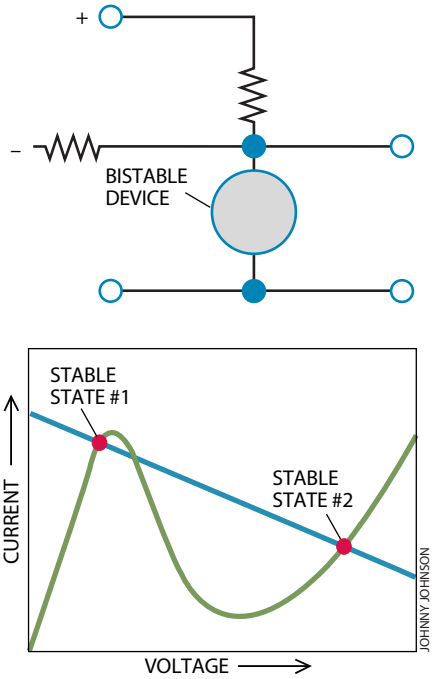
Miniaturization may at first glance appear to involve manipulating just the width and breadth of a device, but depth matters as well. Sometimes the third dimension can be a valuable resource, as when engineers sink capacitors edgewise into a chip to conserve space on the surface. At other times, the third dimension can constrain design. Chip designers must worry about the aspect ratio—that is, the relation of depth to surface area. The devices and connections on chips are built up in the silicon and on the surface as a series of layers resembling a sandwich. Part of the art of making devices smaller comes from using more layers. But the more layers there are, the more carefully controlled each must be, because each is affected by what is beneath it. The number of layers is limited by the costs of better control and more connections between layers.

The formulas that are used to design large devices cannot be used for the tiny transistors now being made in laboratories. Designers need to account for exotic new phenomena that appear in such extremely small devices. Because the effects cannot be accurately treated by purely analytic methods, the designers must have recourse to computer models that are able to simulate the motion of electrons in a device.

A computer follows a single electron through a device, keeping track of its position as time is increased in small steps. Physical theory and experimental information are used to calculate the probability of the various events that are possible. The computer uses a table for the probabilities, stored in its memory, and a random number generator to simulate the occurrence of these events. For example, an electron is accelerated by an electrical field, and the direction of its motion might be changed by a collision with an impurity. Adding the results of thousands of electrons modeled in this fashion gives a picture of the response of the device.

Consider the seemingly trivial question of how to represent the motion of an electron within an electrical field. When path lengths were comparatively long, an electron quickly accelerated to the point at which collisions robbed it of energy as fast as the field supplied new energy. The particle therefore spent most of its time at a constant velocity, which can be modeled by a simple, linear equation. When path lengths became shorter, the electron no longer had time to reach a stable velocity. The particles now accelerate all the time, and the equations must account for that complication.

If such difficulties can arise in modeling a well-understood phenomenon, what lies ahead as designers probe the murky physics of the ultrasmall? Simulations can be no better than the models that physicists make of events that happen in small spaces during short periods. To refine these models, researchers need to carry out experiments on femtosecond timescales.



BISTABLE CIRCUIT does the transistor's job by exploiting nonlinear effects. A device such as a tunnel diode is placed at the junction of two main electrodes and a minor one (*top*). If the minor electrode injects some extra current, the circuit will move from one stable state to the other (*bottom*). Such devices are impractical because they cannot tolerate much variation in signal strength.

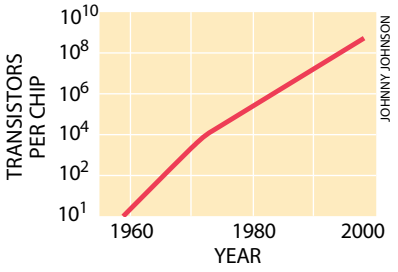
Remaining Unknowns

Expanded knowledge of solid-state physics is required, because as chips grow more complex they require more fabrication steps, and each step can influence the next. For instance, when doping atoms are introduced into a crystal, they tend to attract, repel or otherwise affect the motion of other dopants. Such effects of dopants on other dopants are not well understood; further experiments and theoretical investigations are therefore needed. Chemical reactions that take place on the surface of a silicon crystal demand a supply of silicon atoms, a kind of fluid flow within the solid lattice; does such motion carry other constituents along with it? These questions did not concern designers of earlier generations of chips, because existing transistors were then large enough to swamp such ultramicroscopic tendencies.

Copyright 1997 Scientific American, Inc.

The Shrinking Transistor

Miniaturization is manifest in this comparison between an electromechanical switch, circa 1957 (*background*), and a chip containing 16 million bits of memory (*foreground*). Progress appears in these snapshots (*below, left*): Bell Laboratories's first transistor; canned transistors; salt-size transistors; a 2,000-bit chip; a board with 185,000 circuits and 2.3 megabits of memory; and a 64-megabit memory chip. —*R.W.K.*

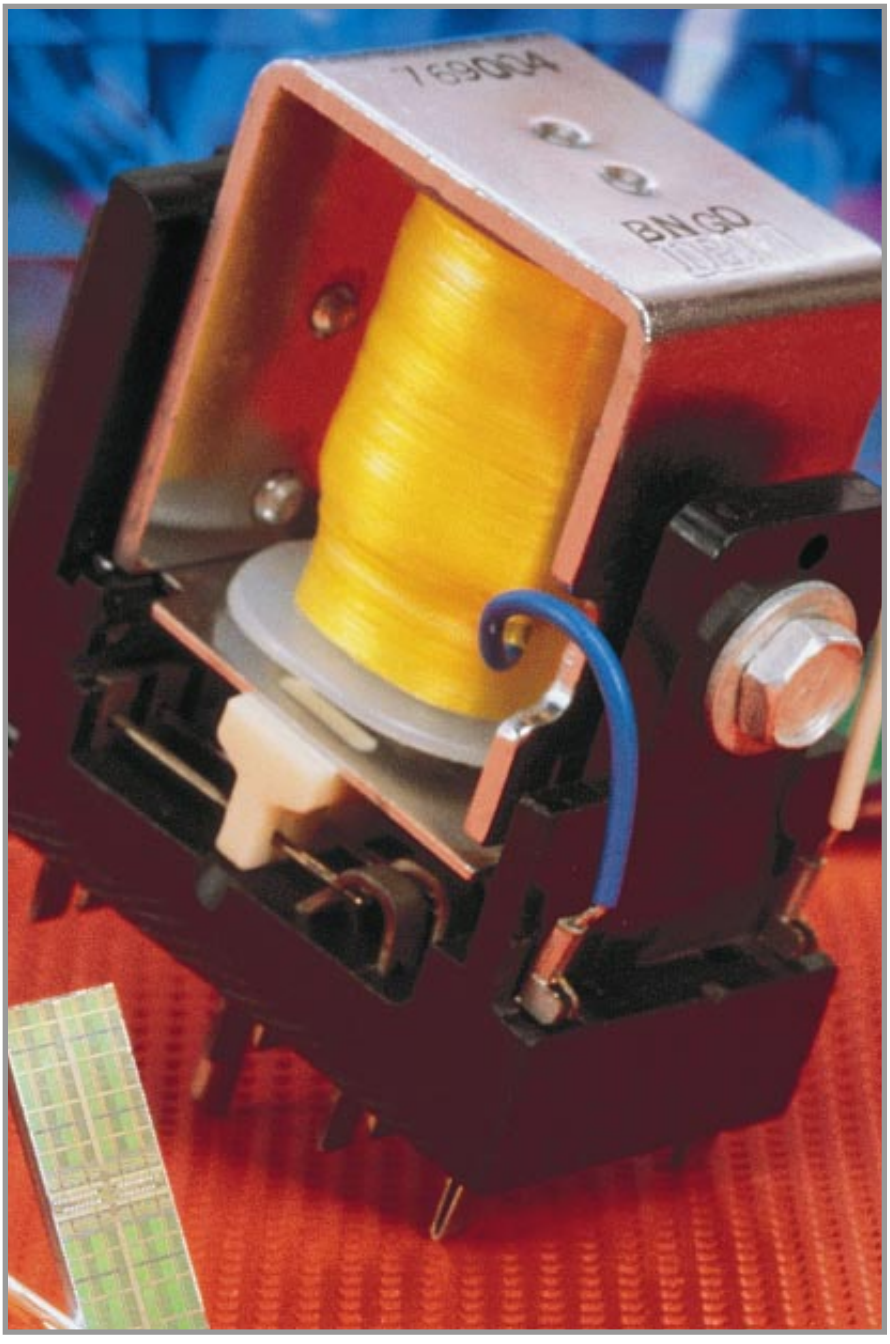


The first transistor
1948

Early commercial transistors
1958

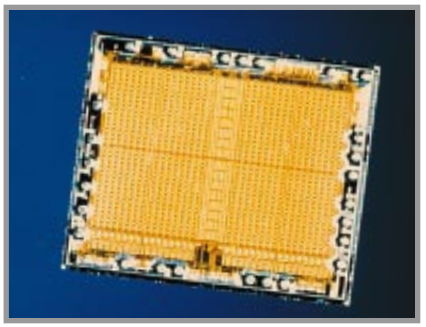
IBM CORPORATION (*photographs*)

Salt-size transistors
1964



Switches, now and then

Copyright 1997 Scientific American, Inc.



Early integrated circuit
1973

Circuit assembly
1985

Dynamic random-access memory chip
1997

The prospect of roadblocks aside, the transistor has only itself to blame for speculation about alternative technologies. Its extraordinary success in the 1950s stimulated an explosive development of solid-state physics. In the course of the work, investigators discovered many other phenomena, which in turn suggested a host of ideas for electronic devices. Several of these lines of research produced respectable bodies of new engineering knowledge but none that led to anything capable of even finding a niche in information processing.

Some workers have argued that the transistor owes its preeminence to having been the first off the block. Because of that head start, semiconductors have been the center of research, a position that guarantees them a margin of technological superiority that no rival can match. Yet I believe the transistor has intrinsic virtues that, in and of themselves, could probably preserve its dominant role for years to come.

I participated, as a minor player, in some of the efforts to build alternative switches, the repeated failures of which made me wonder what was missing. Of course, quite a few new fabrication methods had to be developed to implement a novel device concept. But even though these could be mastered, it was difficult to get a large collection of components to work together.

What gave the transistor its initial, sudden success? One difference stood out: the transistor, like the vacuum tube before it, has large gain. That is, it is capable of vastly amplifying signals of the kind processed in existing circuits, so that a small variation in input can produce a large variation in output. Gain makes it possible to preserve the integrity of a signal as it passes through many switches.

Rivals to the transistor may have been equally easy to miniaturize, but they exhibited far less gain. Take, for instance, bistable devices [*see illustration on page 49*], which perform logic functions by moving between two stable states that are separated by an unstable transition. Researchers have produced such a transition by designing circuits having a range of values in which current declines as voltage increases. Any slight disturbance, such as that obtained by injecting extra current through the device, will switch the circuit between its two stable states.

Because this slight input can bring about large changes in the current and voltages, there is a sense in which gain is achieved. Yet the gain is far less useful than that provided by an ordinary transistor because it operates within rather narrow tolerances. A bistable switch thus performs deceptively well in the laboratory, where it is possible to fine-tune the circuit so it stays near enough to the crossover point. A collection of such switches, however, does not lend itself to such painstaking adjustments. Because not all the circuits will work, no complex device can be based on their operation. Negative resistance therefore plays no role in practical data processing.

The same difficulty has plagued the development of nonlinear optical devices, in which the intensity of optical beams replaces the currents and voltages of electrical circuits. Here, too, the operation depends on fine-tuning the system so that a small input will upset a delicate balance. (Such switches have occasionally been termed optical transistors, a label that misconstrues the principles of transistor action.)

Optical switches face a problem even more fundamental. Light, unlike electricity, hardly interacts with light, yet the interaction of signals is essential for logic functions. Optical signals must therefore be converted into electrical ones in a semiconductor. The voltage thus produced changes the optical response of another material, thereby modulating a beam of light.

Useful Interference

Another proposed switch, sometimes called a quantum interference device, depends on the interference of waves. In the most familiar case, that of electromagnetic radiation, or light, one wave is divided into two components. The components begin oscillating in phase—that is, their peaks and troughs vibrate in tandem. If the components follow routes of different lengths before reuniting, the phase relation between their waveforms will be changed. Consequently, the peaks and troughs either cancel or reinforce one another, producing a pattern of bright and dark fringes. The displacement of the fringes measures the relative phase of the system.

Electrons also possess a wave nature and can be made to interfere. If the two components of a wave move at equal speeds over similar paths to a rendezvous, they will reconstitute the original wave; if they move at different speeds, they will interfere. One can manipulate

Copyright 1997 Scientific American, Inc.

IBM CORPORATION

IMMENSE AND DENSE: this active-matrix liquid-crystal panel shows that today's electronic structures can achieve great complexity over large areas. Each liquid-crystal pixel is controlled by its own transistor, providing extraordinary resolution.

the velocity of one wave by applying a tiny electrical field to its pathway. The correct field strength will cause the waves to cancel so that no current can flow through the device.

At first sight, this action duplicates a field-effect transistor, which uses an electrical field to control a current through a semiconductor. In an interference device, however, conditions must be just right: if the applied voltage is too high or too low, there will be some current. This sensitivity means that an interference device will not restore the binary nature of a degraded input signal but will add its own measure of noise. Data passing from one such device to another will quickly degenerate into nothingness.

The Only Game in Town

The lack of real rivals means that the future of digital electronics must be sought in the transistor. The search begins anew with each voyage into a smaller scale or a different material. The latest reexamination was occasioned by the introduction of new semiconductor materials, such as gallium arsenide and related compounds, several of which may even be incorporated to achieve some desired characteristic in a single device. These combinations may be used to produce what are called heterojunctions, in which crystalline lattices of different energy gaps meet. Lattices may mesh imperfectly, creating atomic-scale defects, or they may stretch to one another, creating an elastic strain. Either defects or strain can produce electrical side effects.

These combinations complicate the physics but at the same time provide a variable that may be useful in surmounting the many design problems that miniaturization creates. For instance, the dopants that supply electrons to a semiconductor also slow the electrons. To reduce this slowing effect, one can alternate layers of two semiconductors in which electrons have differing energies. The dopants are placed in the high-energy semiconductor, but the electrons they donate immediately fall into the lower-energy layers, far from the impurities.

What, one may ask, would one want with a technology that can etch a million transistors into a grain of sand or put a supercomputer in a shirt pocket? The answer goes beyond computational power to the things such power can buy in the emerging information economy. It has only recently been taken for granted that anyone with a personal computer and a modem can search 1,000 newspapers for references to anything that comes to mind, from kiwifruit to quantum physics. Will it soon be possible for every person to carry a copy of the Library of Congress, to model the weather, to weigh alternative business strategies or to checkmate Garry Kasparov? SA

The Author

ROBERT W. KEYES is a research staff member at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y. His interests have focused on semiconductor physics and devices and on the physics of information-processing systems, subjects on which he has written and lectured widely; he has also received eight issued patents. A native of Chicago, Keyes studied at the University of Chicago, where he earned a doctorate in physics. He is an active participant in the programs of the National Research Council, the American Physical Society and the Institute of Electrical and Electronics Engineers.

Further Reading

FIELD-EFFECT TRANSISTORS IN INTEGRATED CIRCUITS. J. T. Wallmark and L. G. Carlstedt. John Wiley & Sons, 1974.

THE PHYSICS OF VLSI SYSTEMS. R. W. Keyes. Addison-Wesley Publishing, 1987.

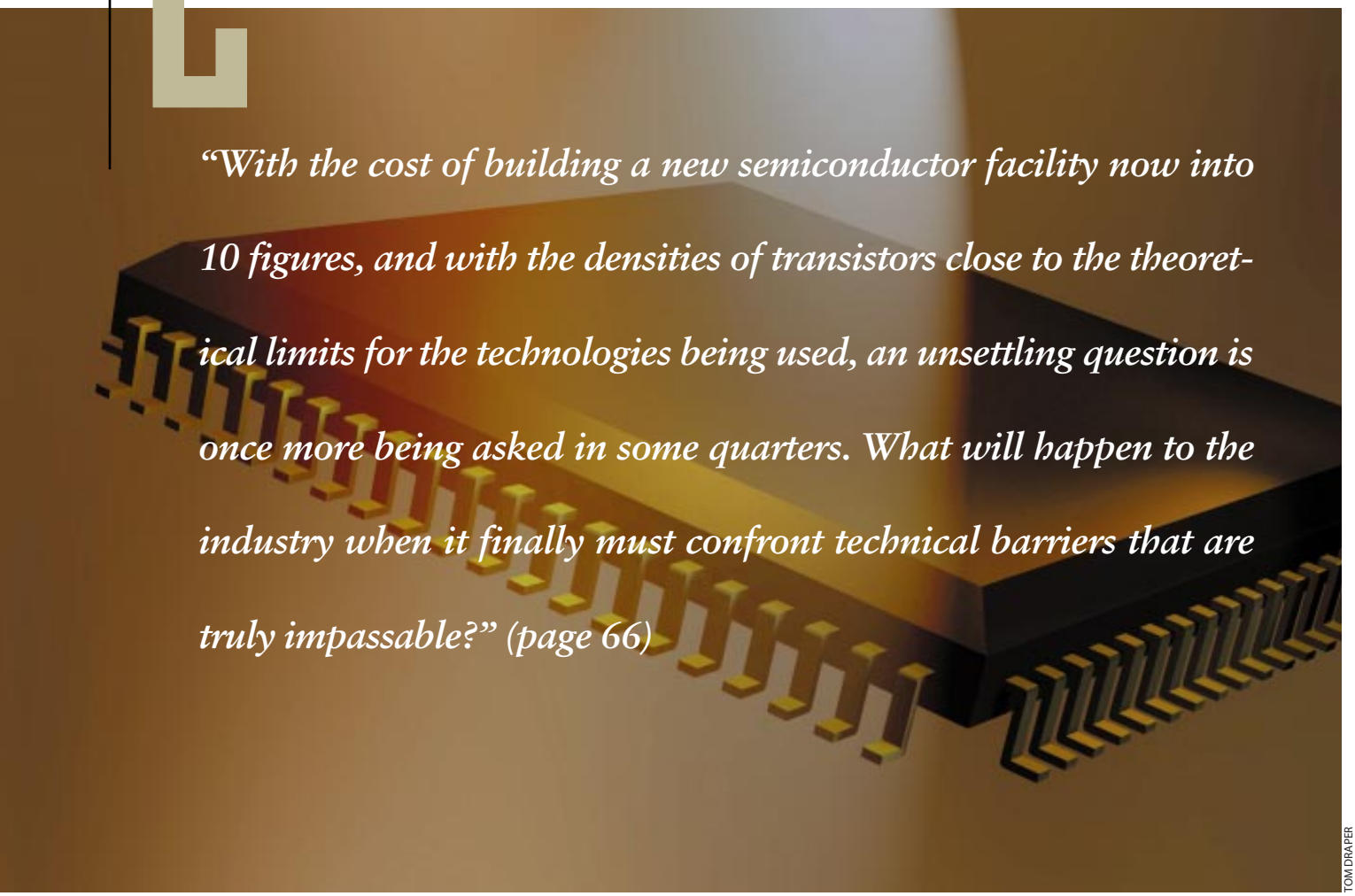
CAN WE SWITCH BY CONTROL OF QUANTUM MECHANICAL TRANSMISSION? Rolf Landauer in *Physics Today,* Vol. 42, No. 10, pages 119–121; October 1989.

LIMITATIONS, INNOVATIONS, AND DEVICES INTO A HALF MICROMETER AND BEYOND. M. Nagata in *IEEE Journal of Solid-State Circuits,* Vol. 27, No. 4, pages 465–472; April 1992.

THE FUTURE OF SOLID-STATE ELECTRONICS. R. W. Keyes in *Physics Today,* Vol. 45, No. 8, pages 42–48; August 1992.

Copyright 1997 Scientific American, Inc.

Integration: The Transistor Meets Mass Production



"With the cost of building a new semiconductor facility now into 10 figures, and with the densities of transistors close to the theoretical limits for the technologies being used, an unsettling question is once more being asked in some quarters. What will happen to the industry when it finally must confront technical barriers that are truly impassable?" *(page 66)*

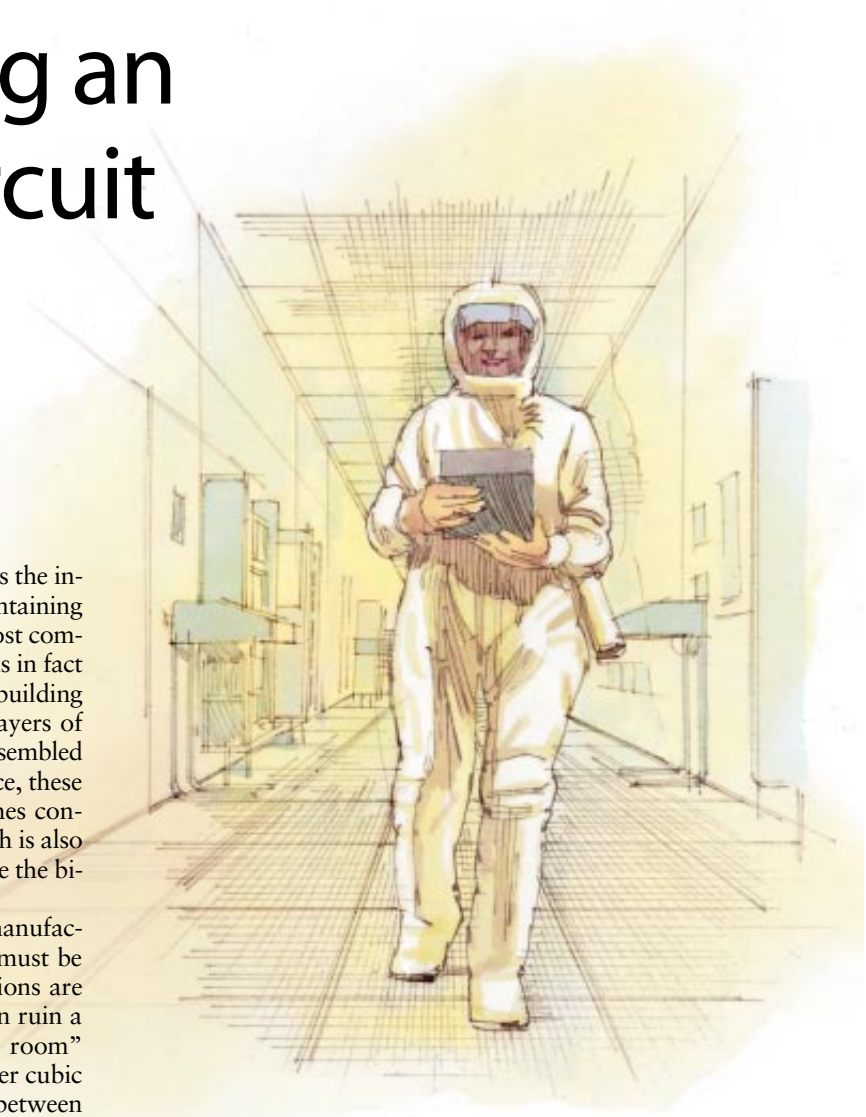
TOM DRAPER

Copyright 1997 Scientific American, Inc.

Tiny silicon chips make modern digital technology possible. Here's how the chips themselves are made

From Sand to Silicon: Manufacturing an Integrated Circuit

by Craig R. Barrett



The fundamental device of the digital world is the integrated circuit, a small square of silicon containing millions of transistors. It is probably the most complex of man-made products. Although it looks flat, it is in fact a three-dimensional structure made by painstakingly building up on the silicon base several microscopically thin layers of materials that both insulate and conduct electricity. Assembled according to a pattern carefully worked out in advance, these layers form the transistors, which function as switches controlling the flow of electricity through the circuit, which is also known as a chip. "On" and "off" switches manipulate the binary code that is at the core of what a computer does.

Building a chip typically requires several hundred manufacturing steps that take weeks to complete. Each step must be executed perfectly if the chip is to work. The conditions are demanding. For example, because a speck of dust can ruin a chip, the manufacturing has to be done in a "clean room" containing less than one submicron particle of dust per cubic foot of air (in contrast, the average living room has between 100,000 and one million particles per cubic foot of air). Much of the equipment needed for making chips embodies the highest of high technology, with the result that chip factories—which cost between $1 billion and $2 billion for a state-of-the-art facility—are among the costliest of manufacturing plants.

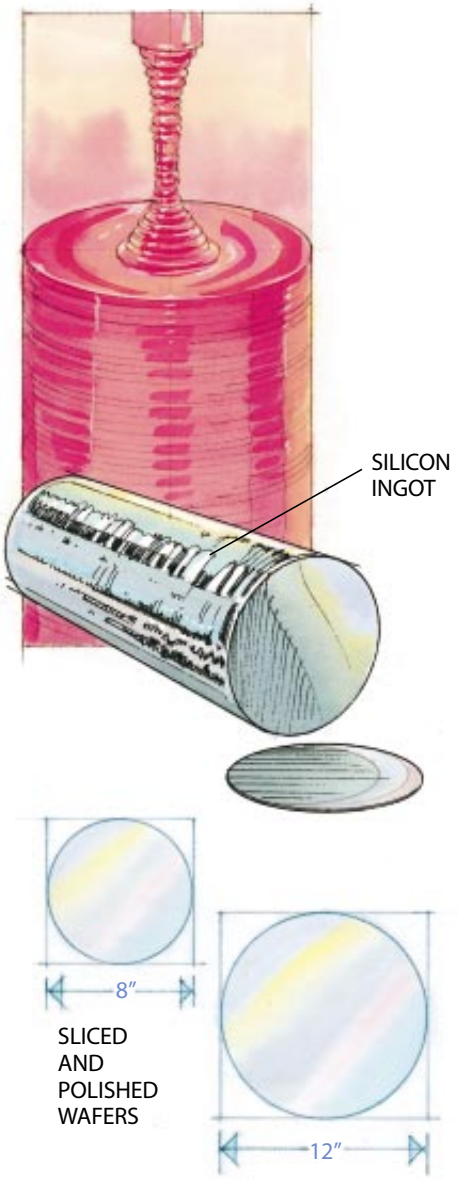
A basic technology of chipmaking is the "planar" process devised in 1957 by Jean Hoerni of Fairchild Semiconductor. It provided a means of creating a layered structure on the silicon base of a chip. This technology was pivotal in Robert N. Noyce's development of the integrated circuit in 1958. (Noyce later became co-founder with Gordon E. Moore of Intel Corporation, the company that invented the microprocessor and has become the world's leading supplier of semiconductor chips. An article about Moore appears on page 62.) Bridging the gap between the transistor and the integrated circuit, the planar technology opened the way to the manufacturing process that now produces chips. The hundreds of individual steps in that process can be grouped into a few basic operations.

CRAIG R. BARRETT is president and chief operating officer of Intel Corporation.

Copyright 1997 Scientific American, Inc.

Chip Design

The first operation is the design of the chip. When tens of millions of transistors are to be built on a square of silicon about the size of a child's fingernail, the placing and interconnections of the transistors must be meticulously worked out. Each transistor must be designed for its intended function, and groups of transistors are combined to create circuit elements such as inverters, adders and decoders. The designer must also take into account the intended purpose of the chip. A processor chip carries out instructions in a computer, and a memory chip stores data. The two types of chips differ somewhat in structure. Because of the complexity of today's chips, the design work is done by computer, although engineers often print out an enlarged diagram of a chip's structure to examine it in detail (*above*).



The Silicon Crystal

The base material for building an integrated circuit is a silicon crystal. Silicon, the most abundant element on the earth except for oxygen, is the principal ingredient of beach sand. Silicon is a natural semiconductor, which means that it can be altered to be either an insulator or a conductor. Insulators, such as glass, block the passage of electricity; conductors, such as copper, let electricity pass through. To make a silicon crystal, raw silicon obtained from quartz rock is treated with chemicals that remove contaminants until what remains is almost 100 percent silicon. This purified silicon is melted and then formed into cylindrical single crystals called ingots (*left, top*). The ingots are sliced into wafers about 0.725 millimeter (0.03 inch) thick. In a step called planarization they are polished with a slurry until they have a flawless, mirror-smooth surface. At present, most of the wafers are 200 millimeters (eight inches) in diameter, but the industry is moving toward achieving a standard diameter of 300 millimeters (12 inches) by 1999 (*left, bottom*). Because a single wafer yields hundreds of chips, bigger wafers mean that more chips can be made at one time, holding down the cost per chip.

DRAWINGS BY BARRY ROSS, AFTER INTEL AND SEMICONDUCTOR EQUIPMENT AND MATERIALS INTERNATIONAL

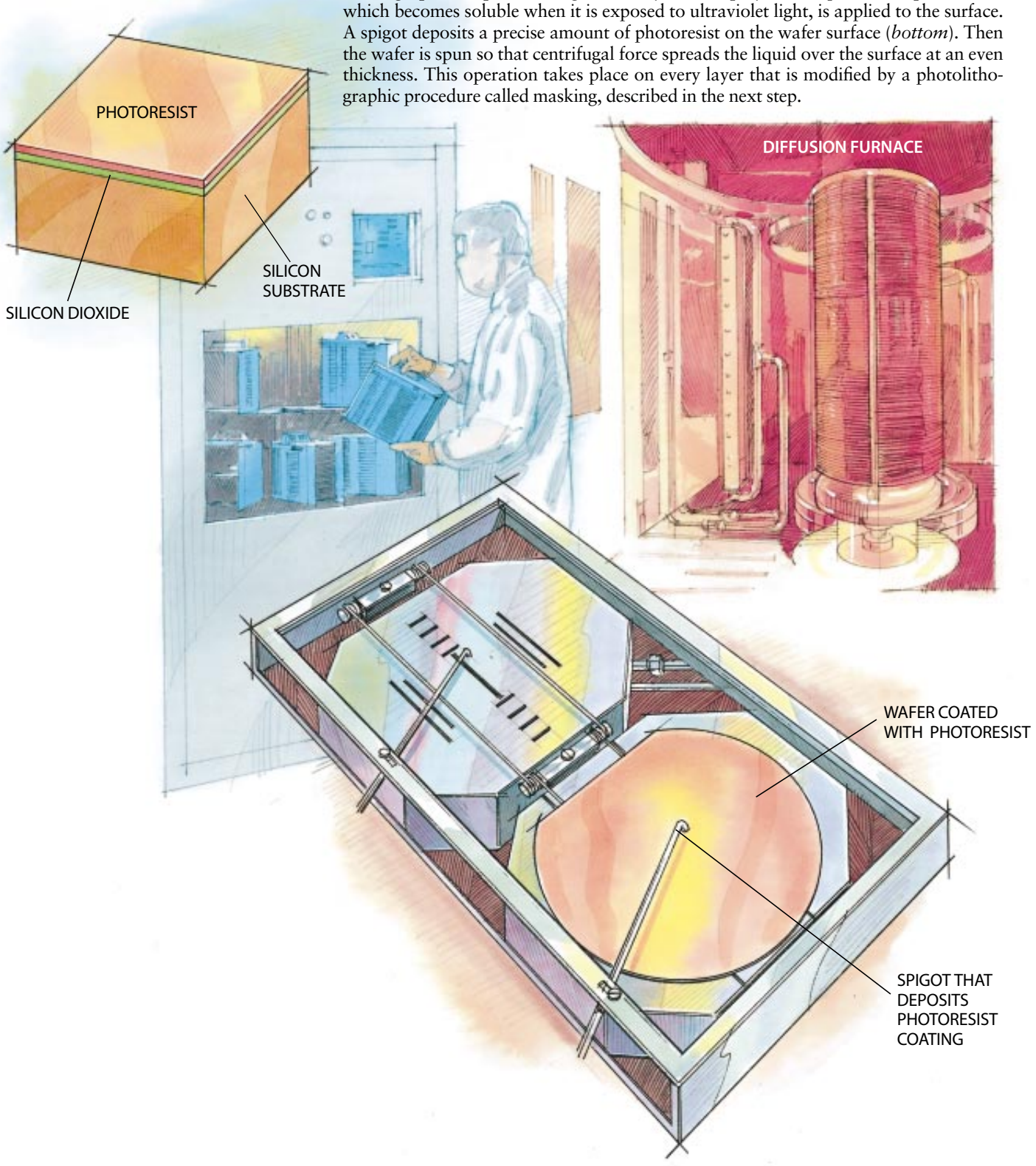
Copyright 1997 Scientific American, Inc.

The First Layers

With the wafer prepared, the process of building the chip's circuitry begins. Making the transistors and their interconnections entails several different basic steps that are repeated many times. The most complex chips made today consist of 20 or more layers and may require several hundred separate processing steps to build them up one by one.

The first layer is silicon dioxide, which does not conduct electricity and therefore serves as an insulator. It is created by putting the wafers into a diffusion furnace (*top right*)—essentially an oven at high temperature where a thin layer of oxide is grown on the wafer surface.

Removed from the furnace, the wafer is now ready for its first patterning, or photolithographic, step. A coating of a fairly viscous polymeric liquid called photoresist, which becomes soluble when it is exposed to ultraviolet light, is applied to the surface. A spigot deposits a precise amount of photoresist on the wafer surface (*bottom*). Then the wafer is spun so that centrifugal force spreads the liquid over the surface at an even thickness. This operation takes place on every layer that is modified by a photolithographic procedure called masking, described in the next step.

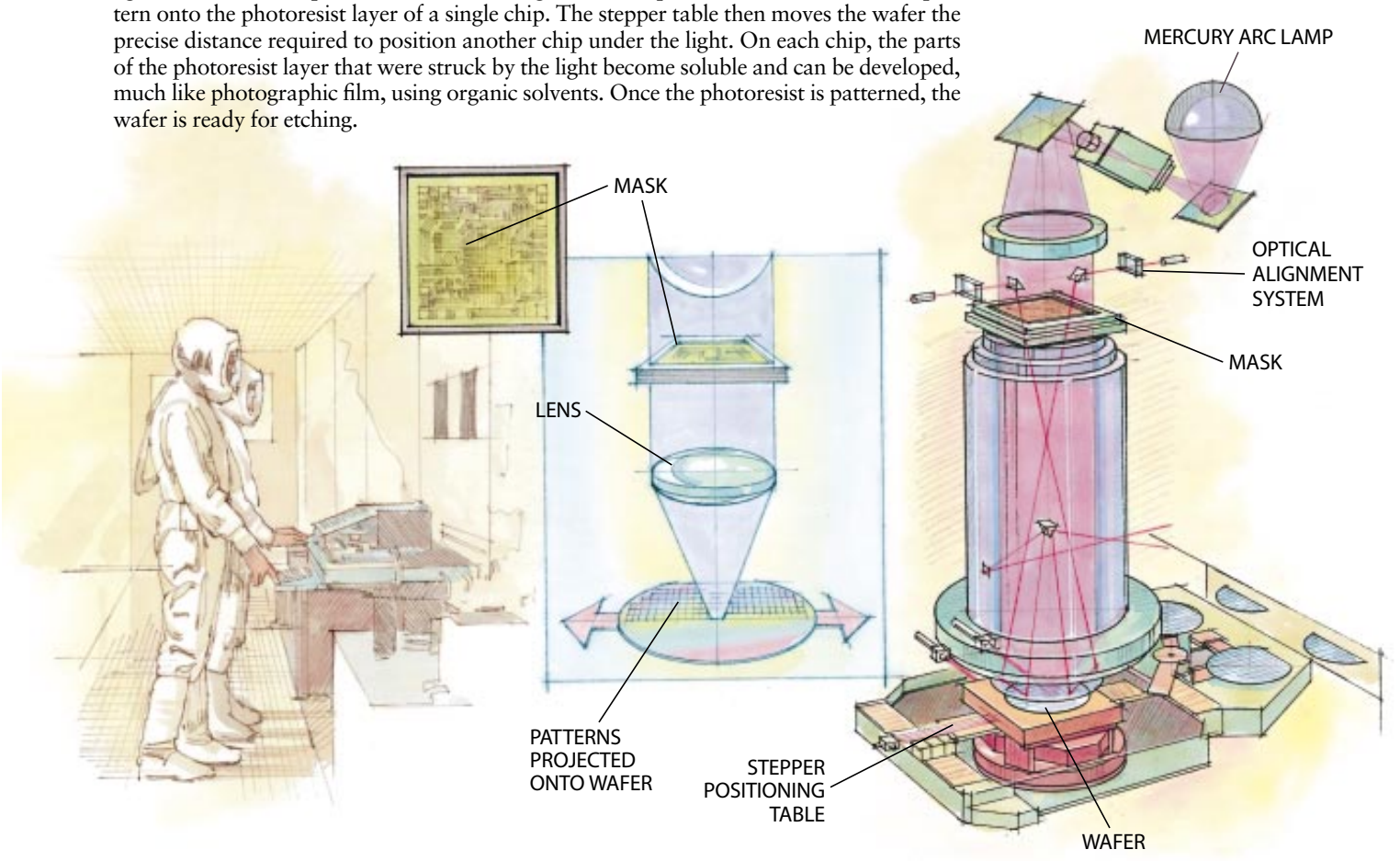


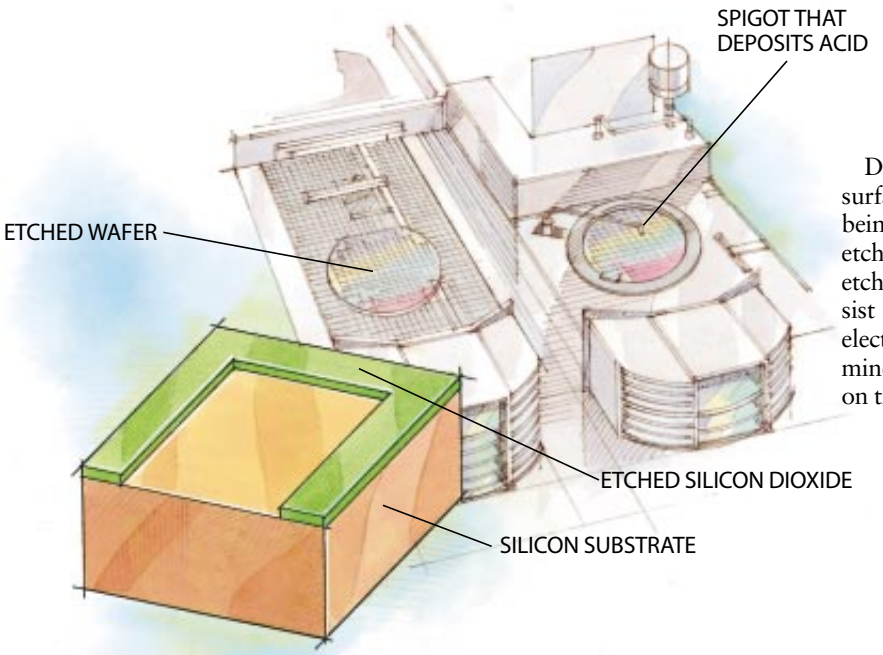
Copyright 1997 Scientific American, Inc.

Masking

A mask is the device through which ultraviolet light shines to define the circuit pattern on each layer of a chip. Because the pattern is intricate and must be positioned precisely on the chip, the arrangement of opaque and transparent spaces on a mask must be done carefully during a chip's design stage.

The mask image is transferred to the wafer using a computer-controlled machine known as a stepper. It has a sophisticated lens system (*below*) to reduce the pattern on the mask to the microscopic dimensions of the chip's circuitry, requiring resolution as small as 0.25 micron. The wafer is held in place on a positioning table below the lens system. Ultraviolet light from an arc lamp or a laser shines through the clear spaces of the mask's intricate pattern onto the photoresist layer of a single chip. The stepper table then moves the wafer the precise distance required to position another chip under the light. On each chip, the parts of the photoresist layer that were struck by the light become soluble and can be developed, much like photographic film, using organic solvents. Once the photoresist is patterned, the wafer is ready for etching.





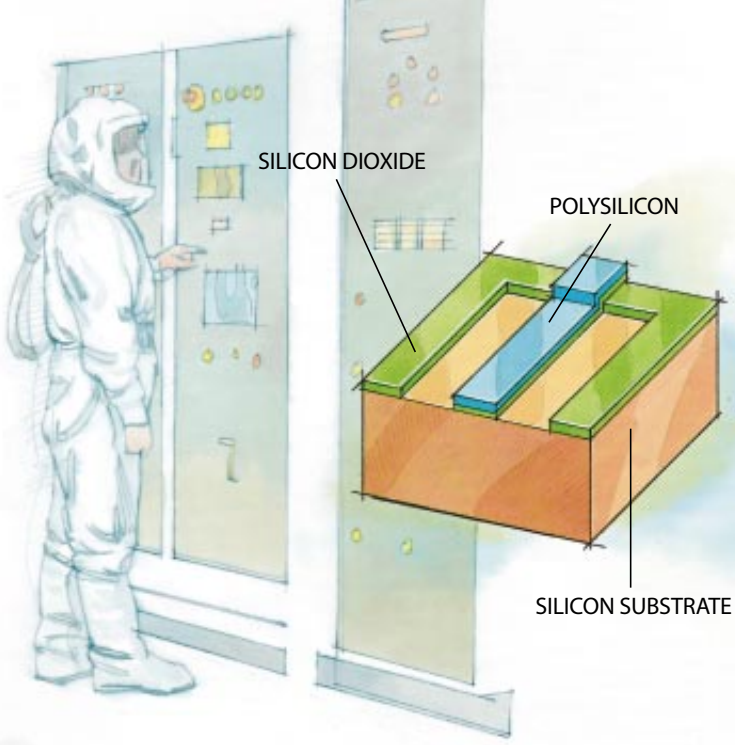
Etching

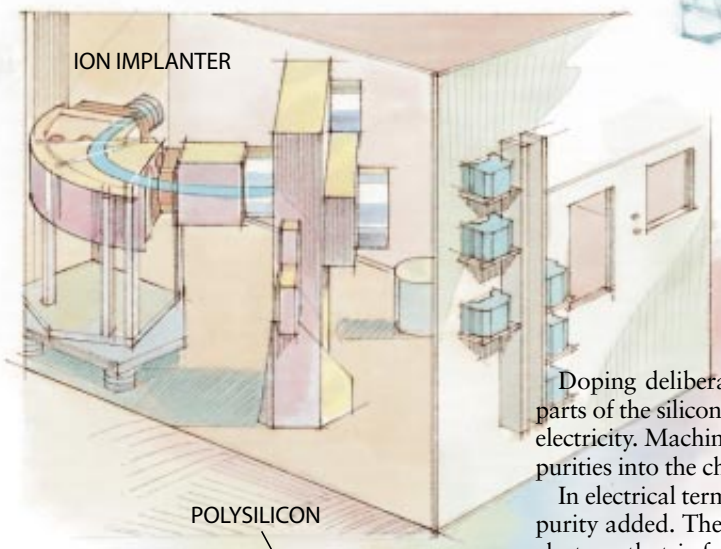
During this step, photoresist remaining on the surface protects parts of the underlying layer from being removed by the acids or reactive gases used to etch the pattern on the surface of the wafer. After etching is complete, the protective layer of photoresist is removed to reveal electrically conducting or electrically insulating segments in the pattern determined by the mask (*left*). Each additional layer put on the chip has a distinctive pattern of this kind.

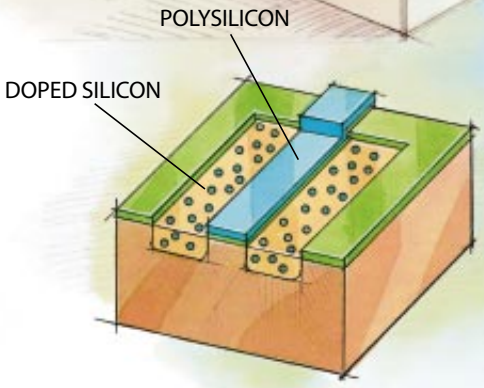
Copyright 1997 Scientific American, Inc.

Adding Layers

Further masking and etching steps deposit patterns of additional materials on the chip. These materials include polysilicon as well as various oxides and metal conductors such as aluminum and tungsten. To prevent the formation of undesired compounds during subsequent steps, other materials known as diffusion barriers can also be added. On each layer of material, masking and etching create a unique pattern of conducting and nonconducting areas (*right*). Together these patterns aligned on top of one another form the chip's circuitry in a three-dimensional structure. But the circuitry needs fine-tuning to work properly. The tuning is provided by doping.







Doping

Doping deliberately adds chemical impurities, such as boron or arsenic, to parts of the silicon wafer to alter the way the silicon in each doped area conducts electricity. Machines called ion implanters (*left*) are often used to inject these impurities into the chip.

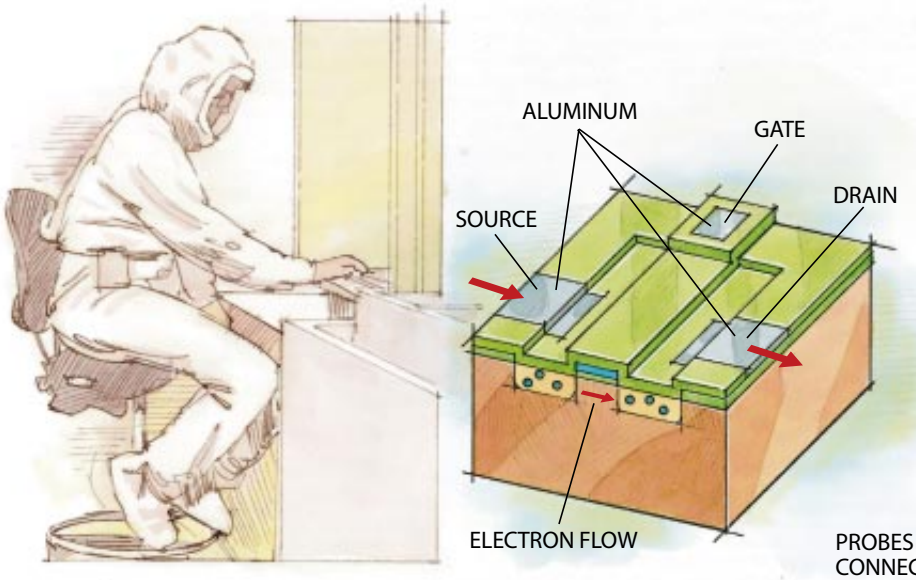
In electrical terms, silicon can be either *n*-type or *p*-type, depending on the impurity added. The atoms in the doping material in *n*-type silicon have an extra electron that is free to move. Some of the doping atoms in *p*-type silicon are short an electron and so constitute what is called a hole. Where the two types adjoin, the extra electrons can flow from the *n*-type to the *p*-type to fill the holes.

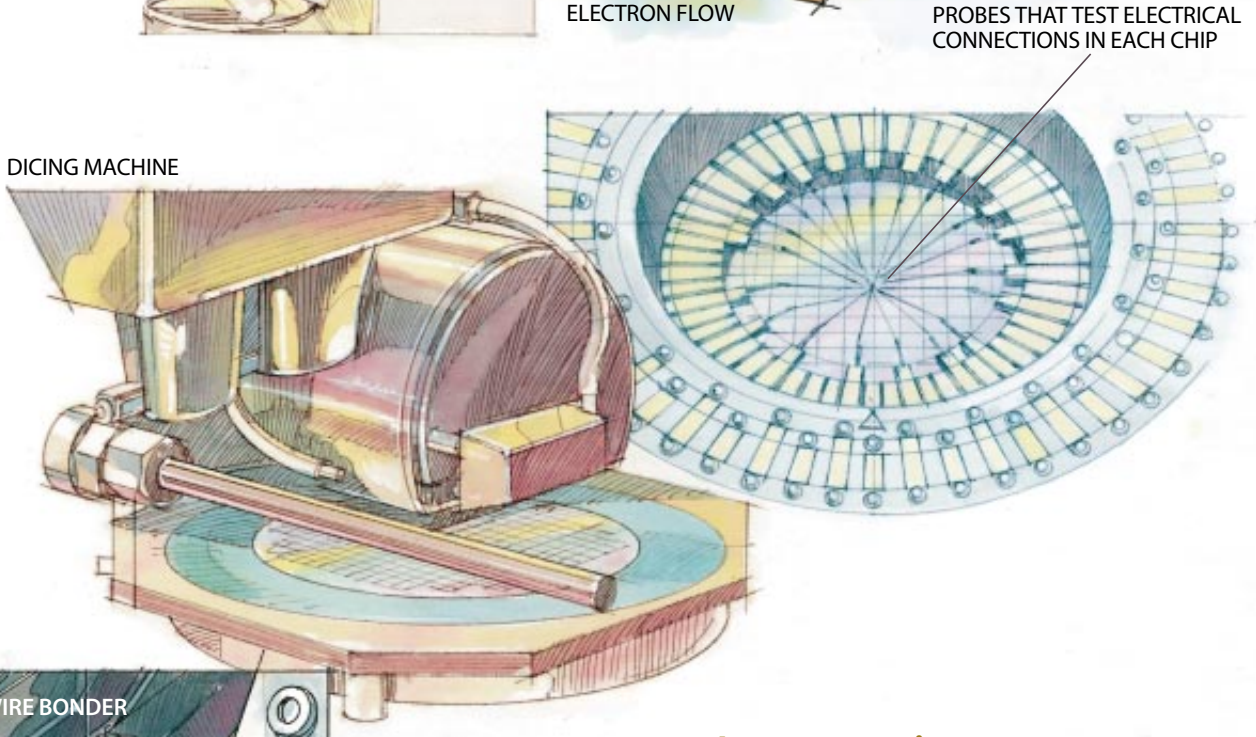
This flow of electrons does not continue indefinitely. Eventually the positively charged ions left behind on the *n*-type side and the negatively charged ions on the *p*-type side together create an electrical force that prevents any further net flow of electrons from the *n*-type to the *p*-type region.

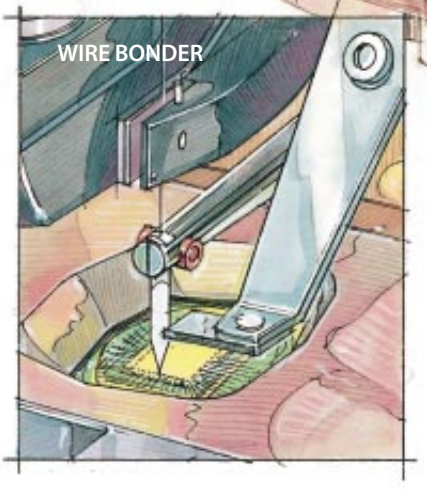
The material at the base of the chip is *p*-type silicon. One of the etching steps in the manufacture of a chip removes parts of the polysilicon and silicon dioxide layers put on the pure silicon base earlier, thus laying bare two strips of *p*-type silicon. Separating them is a strip that still bears its layer of conducting polysilicon; it is the transistor's "gate." The doping material now applied to the two strips of *p*-type silicon transforms them into *n*-type silicon. A positive charge applied to the gate attracts electrons below the gate in the transistor's silicon base. These electrons create a channel between one *n*-type strip (the source) and the other (the drain). If a positive voltage is applied to the drain, current will flow from source to drain. In this mode, the transistor is "on." A negative charge at the gate depletes the channel of electrons, thereby preventing the flow of current between source and drain. Now the transistor is "off." It is by means of switching on and off that a transistor represents the arrays of 1 and 0 that constitute the binary code, the language of computers.

Done many times in many layers, these operations provide the chip with its multitude of transistors. But just as provision must be made to run electrical wires and plumbing pipes between floors of a building, provision must be made in chips for interconnecting the transistors so they form an integrated circuit.

Copyright 1997 Scientific American, Inc.







Interconnections

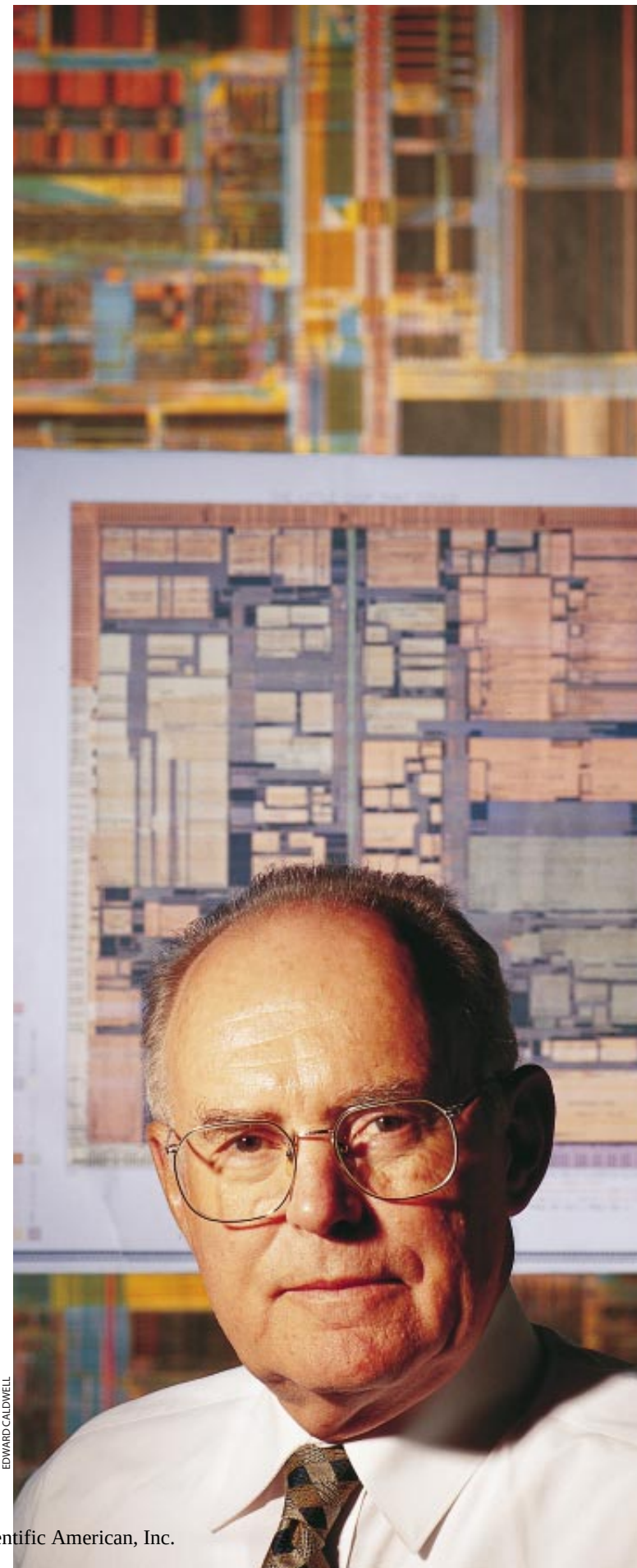
This final step begins with further masking and etching operations that open a thin layer of electrical contacts between layers of the chip. Then aluminum is deposited and patterned using photolithography to create a form of wiring that links all the chip's transistors (*top*). Aluminum is chosen for this application because it makes good electrical contact with silicon and also bonds well to silicon dioxide.

This step completes the processing of the wafer. Now the individual chips are tested to ensure that all their electrical connections work using tiny electrical probes (*above right*). Next, a machine called a dicer cuts up the wafer into individual chips (*above left*), and the good chips are separated from the bad. The good chips—usually most of the wafer's crop—are mounted onto packaging units with metal leads. Wire bonders (*left*) then attach these metal leads to the chips. The electrical contacts between the chip's surface and the leads are made with tiny gold or aluminum wires about 0.025 millimeter (0.001 inch) in diameter. Once the packaging process is complete, the finished chips are sent to do their digital work. SA

Copyright 1997 Scientific American, Inc.

The Law of More

by W. Wayt Gibbs, *staff writer*



EDWARD CALDWELL

Technologists are given to public displays of unbridled enthusiasm about the prospects of their inventions. So a reader flipping through the 35th anniversary issue of *Electronics* in April 1965 might easily have dismissed an article by Gordon E. Moore, then head of research at Fairchild Semiconductor, pitching the future of his business. Moore observed that the most cost-effective integrated circuits had roughly doubled in complexity each year since 1959; they now contained a whopping 50 transistors per chip. At that rate, he projected, microchips would contain 65,000 components by 1975, at only a modest increase in price. "Integrated circuits," Moore wrote, "will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment."

Technically, Moore was overoptimistic: 65,000-transistor chips did not appear until 1981. But his fundamental insight—that continued geometric growth in the complexity of microelectronics would be not only feasible but also profitable—held true for so long that others began referring to it as Moore's Law. Today, from his vantage point as chairman emeritus of Intel, Moore observes that his prediction "has become a self-fulfilling prophecy. [Chipmakers] know they have to stay on that curve to remain competitive, so they put the effort in to make it happen."

That effort grows with each generation—Intel and its peers now spend about $20 billion a year on research. Moore expects the rule of his law to end within the next decade, coinciding nicely with the twilight of his career. Such good fortune—the kind that tends to smile on the prepared—is a recurrent theme in the history of Moore and the microprocessor.

Even Moore's entry into the semiconductor business was accidental. A year after finishing his doctorate at the California Institute of Technology in 1954, the physical chemist decided to take a job as an inspector of nuclear explosions at Lawrence Livermore National Laboratory. By coincidence, William Shockley, one of the inventors of the transistor, was at the time looking for chemists to work in his semiconductor company and got permission to rifle through Livermore's résumé file. "I had no background whatsoever in semiconductors," Moore recalls. Shockley offered him a job anyway.

"Shockley was a technical genius, but he really didn't understand how people worked very well," Moore says. Within a year he, Robert N. Noyce and several colleagues abandoned Shockley to found a new firm. Fairchild Semiconductor produced the first commercial integrated circuit in 1959 and grew over the next decade into a $150-million business.

Copyright 1997 Scientific American, Inc.

Gordon E. Moore co-founded two high-tech titans but is best known for an eponymous law that may finally be nearing its limit

But soon after it was bought out by a conglomerate, Moore grew restive. In 1968 he and Noyce struck out again on their own.

Fairchild and its competitors were still customizing chips for every system. "The idea we had for Intel," Moore says, was "to make something complex and sell it for all kinds of digital applications": first memory chips, then calculators. "But we were a little late," Moore adds. All the big calculator companies already had partners.

Noyce tracked down a small Japanese start-up named Busicom that had designed the logic for 13 microcircuits to go into its scientific calculators. "To do 13 different complex custom circuits was far beyond what we could tackle," Moore recounts. But after some thought, Intel engineer Ted Hoff concluded that a single, general-purpose chip could perform all 13 functions and more.

And so, out of chance and desperate necessity, the microprocessor was born in 1971. Under Moore's direction, four Intel engineers created, in nine months, a computer on a chip. There was just one problem, Moore admits, with a sheepish grin: "Busicom paid a portion of the development costs and therefore owned the rights to the design."

But fortune was again on Moore's side. Busicom slipped into financial straits. "We essentially gave them back the $65,000 [they had paid Intel] and got the rights to the chips back for all uses. So the Japanese initially owned all the rights to microprocessors but sold them for 65 grand. In retrospect, it was kind of like the purchase of Manhattan" island for $24 in 1626, he laughs.

With Moore as chief executive and later as chairman, Intel rode the wave of Moore's Law for over 25 years. But that wave will begin to break as the costs of cramming more transistors on a slice of silicon overwhelm the benefits [see "Toward 'Point One,'" by Gary Stix, page 74]. "Things we used to do relatively casually to advance the technology now take teams of Ph.D.'s," he says, estimating that 400 engineers worked for four years to produce Intel's latest processors.

Moore predicts that tweaking the lenses, robots and ultraviolet lasers used to etch circuits onto silicon will extract perhaps two more generations of processors, with features 0.18, then 0.13 micron across, from current optical techniques. "Beyond that, life gets very interesting," he notes. "We have three equally unattractive alternatives."

X-rays, with their smaller wavelength, could carve out wires just a handful of atoms across. But blocking such energetic waves requires very thick stencils as tiny as the chip itself. "It is very hard to make the mask perfect enough and then to do the precision alignment," Moore warns. "So while a lot of work continues on x-rays, some of us have lost our enthusiasm for that technology."

A second option is to use electron beams to draw circuit designs line by line onto silicon. But that process is still far too slow for mass production, Moore says: "And as you go to smaller dimensions, the total distance the beam has to travel to make the pattern keeps going up." Experiments with wider beams look promising, however. "Worst case, we will be able to make a layer or two of some very fine structures with an electron beam, then add optically some structures that are not so fine," he wagers.

The smart money, Moore says, is on soft (relatively low frequency) x-rays. "There is still a tremendous amount of engineering involved in making this work," he cautions. "You have to have a reflective mask instead of a transparent mask. You have to have a vacuum system. You have to have new resist [coatings]." But if it succeeds, Moore concludes, soft x-ray lithography "will take us as far as the material will let us go, a long ways from where it is now."

Moore worries as much about the consequences of tinier transistors as about ways to make them. With the rapid increase in chips' complexity and clock speeds, "if you don't do anything else, the power goes up something like 40-fold" every two generations, he points out. "Well, if you start with a 10-watt device [such as the Pentium] and go up 40-fold, the darn thing smokes! We've handled it to date by lowering the voltage. But you can only go so far on that."

To squeeze out another decade of geometric performance growth, chip manufacturers will try various tricks, Moore predicts. "Phase-shift masks [that compensate for the diffraction of laser light] allow you to go to smaller dimensions with a given wavelength." Processors are already five or six layers high; they will thicken still more. And silicon wafers will grow from eight to 12 inches in diameter, enabling greater economies of scale. Until recently, Moore concedes, "I was a skeptic there. I argued that the cost of material was going to be prohibitive." But he failed to foresee advances in crystal-growth techniques.

Moore's vision is also less than clear on what new workaday jobs will require desktop supercomputers with processors that are 10 to 100 times more powerful than today's. "We haven't identified any very general ones yet," he admits. Indeed, by the time computer hardware begins to lag Moore's vision, engineers may find that the barriers to more intelligent, useful machines lie not in physics but in software, which obeys no such law. SA

INTEL CO-FOUNDER Gordon E. Moore sees potholes on the road to desktop supercomputers.

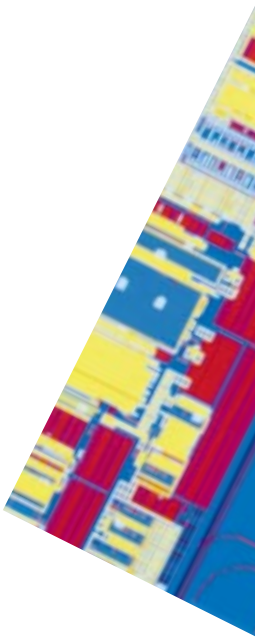
The full text of Scientific American*'s interview with Gordon E. Moore is available at http://www.sciam.com*

Copyright 1997 Scientific American, Inc.

Although the days of runaway growth may be numbered, their passing may force chipmakers to offer more variety

Technology and Economics in the Semiconductor Industry

by G. Dan Hutcheson and Jerry D. Hutcheson



The ability to store and process information in new ways has been essential to humankind's progress. From early Sumerian clay tokens through the Gutenberg printing press, the Dewey decimal system and, eventually, the semiconductor, information storage has been the catalyst for increasingly complex legal, political and societal systems. Modern science, too, is inextricably bound to information processing, with which it exists in a form of symbiosis. Scientific advances have enabled the storage, retrieval and processing of ever more information, which has helped generate the insights needed for further advances.

Over the past few decades, semiconductor electronics has become the driving force in this crucial endeavor, ushering in a remarkable epoch. Integrated circuits made possible the personal computers that have transformed the world of business, as well as the controls that make engines and machines run more cleanly and efficiently and the medical systems that save lives. In so doing, they spawned industries that are able to generate hundreds of billions of dollars in revenues and provide jobs for millions of people. All these benefits, and far too many more to list here, accrue in no small measure from the fact that the semiconductor industry has been able to integrate more and more transistors onto chips, at ever lower costs.

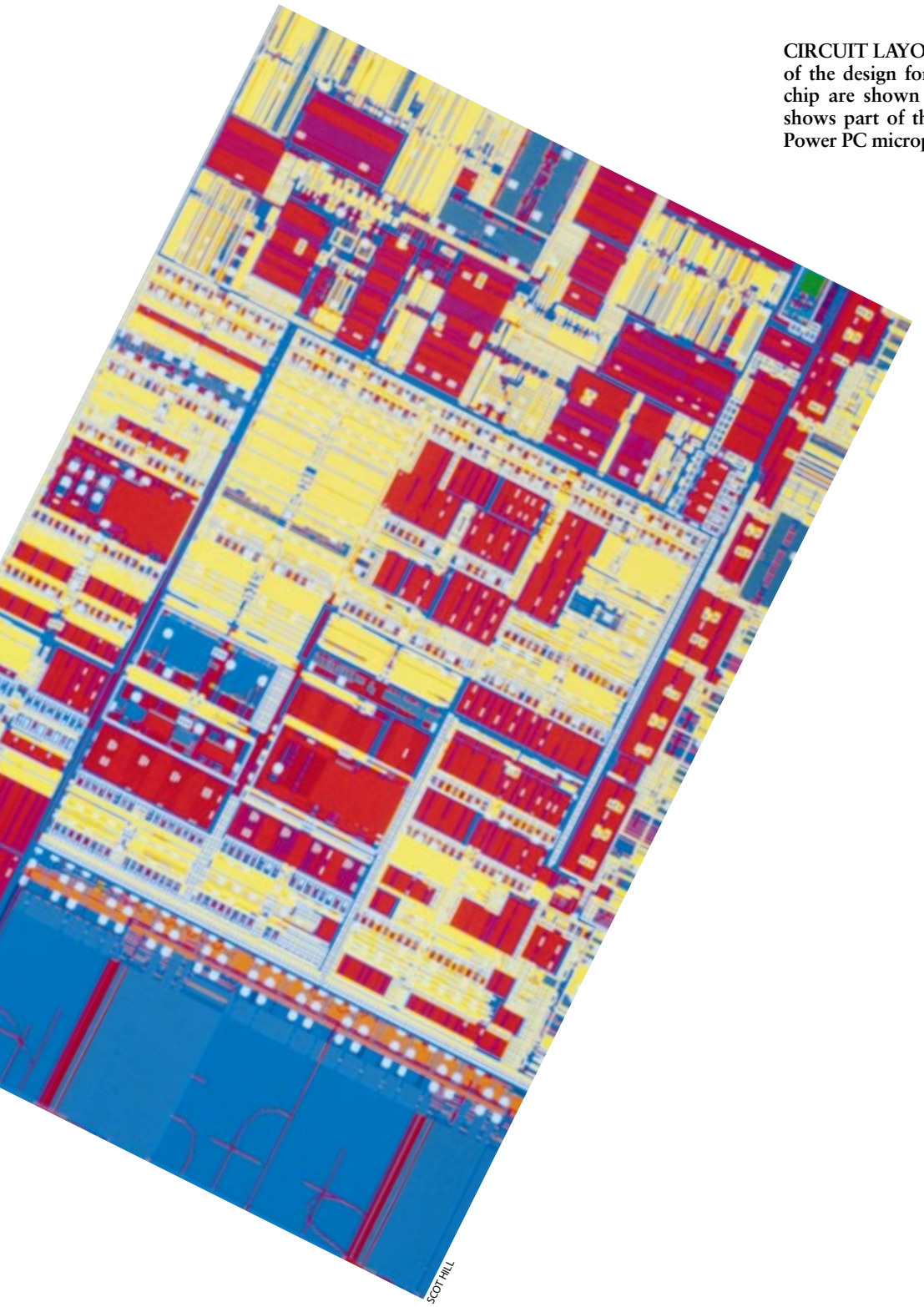
This ability, largely unprecedented in industrial history, is so fundamental in the semiconductor business that it is literally regarded as a law. Nevertheless, from time to time, fears that technical and economic obstacles might soon slow the pace of advances in semiconductor technology have cropped up. Groups of scientists and engineers have often predicted the imminence of so-called showstopping problems, only to see those predictions foiled by the creativity and ingenuity of their peers.

To paraphrase a former U.S. president, here we go again. With the cost of building a new semiconductor facility now into 10 figures, and with the densities of transistors close to the theoretical limits for the technologies being used, an unsettling question is once more being asked in some quarters. What will happen to the industry when it finally must confront technical barriers that are truly impassable?

Moore and More Transistors

In 1965, seven years after the integrated circuit was invented, Gordon Moore observed that the number of transistors that semiconductor makers could put on a chip was doubling every year. Moore, who cofounded Intel Corporation in 1968 and is now an industry sage, correctly predicted that this pace would continue into at least the near future. The phenomenon became known as Moore's Law, and it has had far-reaching implications.

Copyright 1997 Scientific American, Inc.



CIRCUIT LAYOUT helps designers to keep track of the design for a chip. Different layers of the chip are shown in different colors. This image shows part of the layout for one of Motorola's Power PC microprocessors.

himself—integration continued to increase at an astounding rate. True, in the late 1970s, the pace slowed to a doubling of transistors every 18 months. But it has held to this rate ever since, leading to commercial integrated circuits today with more than six million transistors. The electronic components in these chips measure 0.35 micron across. Chips with 10 million or more transistors measuring 0.25 or even 0.16 micron are expected to become commercially available soon.

In stark contrast to what would seem to be implied by the dependable doubling of transistor densities, the route that led to today's chips was anything but smooth. It was more like a harrowing obstacle course that repeatedly required chipmakers to overcome significant limitations in their equipment and production processes. None of those problems turned out to be the dreaded showstopper whose solution would be so costly that it would slow or even halt the pace of advances in semiconductors and, therefore, the growth of the industry. Successive roadblocks, however, have become increasingly imposing, for reasons tied to the underlying technologies of semiconductor manufacturing.

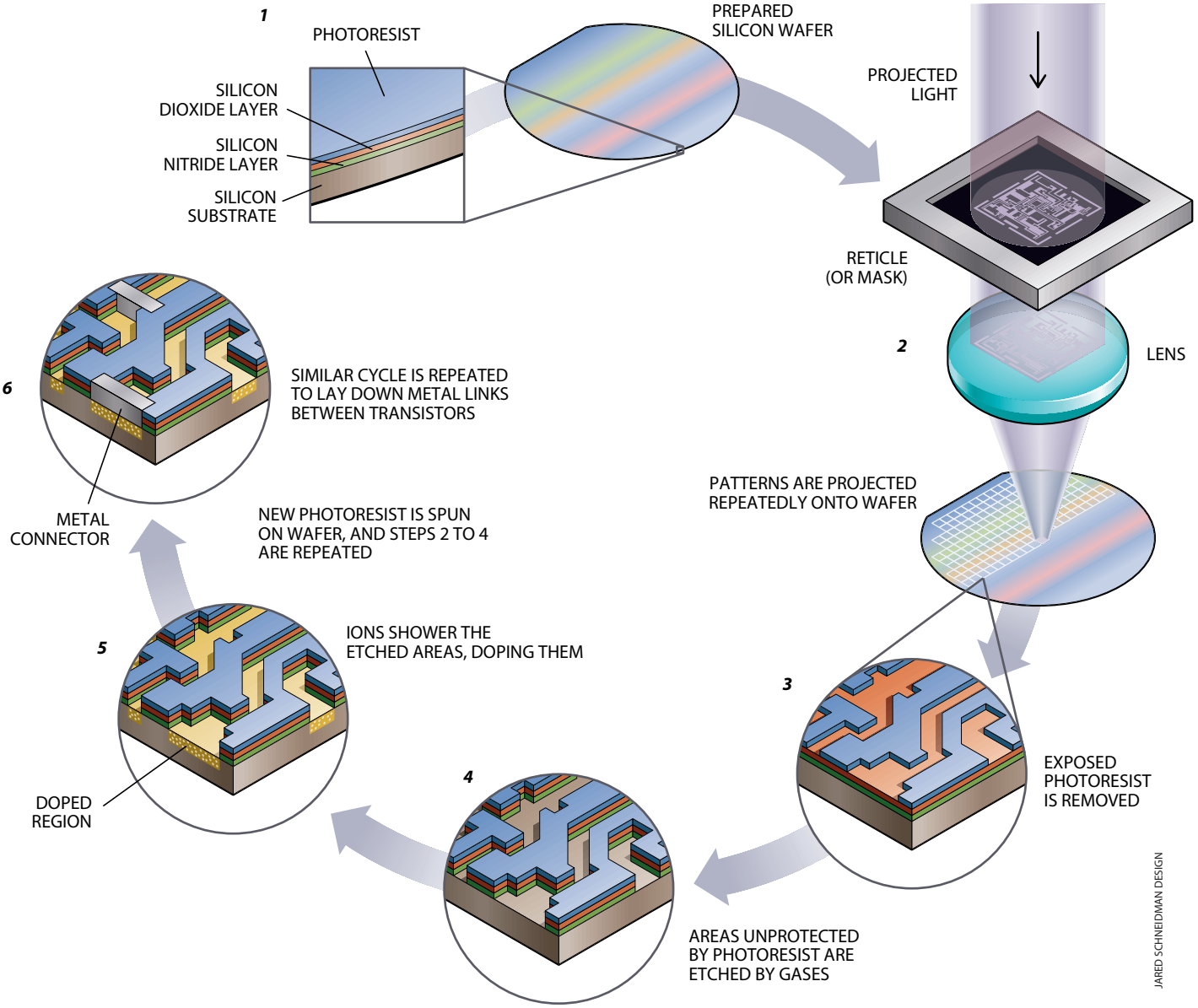
Chips are made by creating and interconnecting transistors to form complex electronic systems on a sliver of silicon. The fabrication process is based on a series of steps, called mask layers, in which films of various materials—some sensitive to light—are placed on the silicon and exposed to light. After these deposition and lithographic procedures, the layers are processed to "etch" the patterns that, when precisely aligned and combined with those on successive layers, produce the transistors and connections. Typically, 200 or more chips are fabricated simultaneously on a thin disk, or wafer, of silicon [*see illustration on next page*].

Because the doublings in density were not accompanied by an increase in cost, the expense per transistor was halved with each doubling. With twice as many transistors, a memory chip can store twice as much data. Higher levels of integration mean greater numbers of functional units can be integrated onto the chip, and more closely spaced devices, such as transistors, can interact with less delay. Thus, the advances gave users increased computing power for the same money, spurring both sales of chips and demand for yet more power.

To the amazement of many experts—including Moore

In the first set of mask layers, insulating oxide films are de-

Copyright 1997 Scientific American, Inc.



CHIP FABRICATION occurs as a cycle of steps carried out as many as 20 times. Many chips are made simultaneously on a silicon wafer, to which has been applied a light-sensitive coating (*1*). Each cycle starts with a different pattern, which is projected repeatedly onto the wafer (*2*). In each place where the image falls, a chip is made. The photosensitive coating is removed (*3*), and the light-exposed areas are etched by gases (*4*). These areas are then showered with ions (or "doped"), creating transistors (*5*). The transistors are then connected as successive cycles add layers of metal and insulator (*6*).

posited to make the transistors. Then a photosensitive coating, called the photoresist, is spun over these films. The photoresist is exposed with a stepper, which is similar to an enlarger used to make photographic prints. Instead of a negative, however, the stepper uses a reticle, or mask, to project a pattern onto the photoresist. After being exposed, the photoresist is developed, which delineates the spaces, known as contact windows, where the different conducting layers interconnect. An etcher then cuts through the oxide film so that electrical contacts to transistors can be made, and at that point, the photoresist is removed.

More sets of mask layers, based on much the same deposition, lithography and etching steps, create the conducting films of metal or polysilicon needed to link transistors. All told, about 19 mask layers are required to make a chip.

The physics underlying these manufacturing steps suggests several potential obstacles to continued technical progress. One follows from Rayleigh's resolution limit, named after John William Strutt, the third Baron of Rayleigh, who won the 1904 Nobel Prize for Physics. According to this limit, the size of the smallest features that can be resolved by an optical system with a circular aperture is proportional to the wavelength of the light source divided by the diameter of the aperture of the objective lens. In other words, the shorter the wavelength and the larger the aperture, the finer the resolution.

The limit is a cardinal law in the semiconductor industry because it can be used to determine the size of the smallest transistors that can be put on a chip. In the lithography of integrated circuits, the most commonly used light source is the mercury lamp. Its most

Copyright 1997 Scientific American, Inc.

useful line spectra for this purpose occur at 436 and 365 nanometers, the so-called mercury g and i lines. The former is visible to the human eye; the latter is just beyond visibility in the ultraviolet. The numerical apertures used range from a low of about 0.28 for run-of-the-mill industrial lenses to a high of about 0.65 for those in leading-edge lithography tools. These values, taken together with other considerations arising from demands of high-volume manufacturing, give a limiting resolution of about 0.54 micron for g-line lenses and about 0.48 for i-line ones.

Until the mid-1980s, it was believed that g-line operation was the practical limit. But one by one, obstacles to i-line operation were eliminated in a manner that well illustrates the complex relations between economics and technology in the industry. Technical barriers were surmounted, and, more important, others were found to be mere by-products of the level of risk the enterprise was willing to tolerate. This history is quite relevant to the situation the industry now finds itself in—close to what appear to be the practical limits of i-line operation.

Must the Show Go On?

One of the impediments to i-line operation, at the time, was the fact that most of the glasses used in lenses are opaque at i-line frequencies. Only about 70 percent of i-line radiation passes through these lenses; the rest is converted to heat, which can distort the image. Many experts believed these factors would necessitate the use of quartz. Even if practical quartz lenses could be made, it was reasoned, verifying the alignment of patterns that could not be seen would be difficult. Glasses were eventually developed that could pass more than 99 percent of i-line radiation, and new technologies were invented to solve the alignment problem.

There are other difficulties, however. Rayleigh's limit also establishes the interval within which the pattern projected by the lens is in focus. Restricted depth of focus can work against resolution limits: the better the resolution, the shallower the depth of focus. For a lens as described above, the depth of focus is about 0.52 micron for the best g-line lenses and about 0.50 for i-line ones. Such shallow depths of focus demand extremely flat wafer surfaces—much flatter than what could be maintained across the diagonal of a large chip with the best available equipment just several years ago.

Skyrocketing costs have once again focused attention on limits in the semiconductor industry.

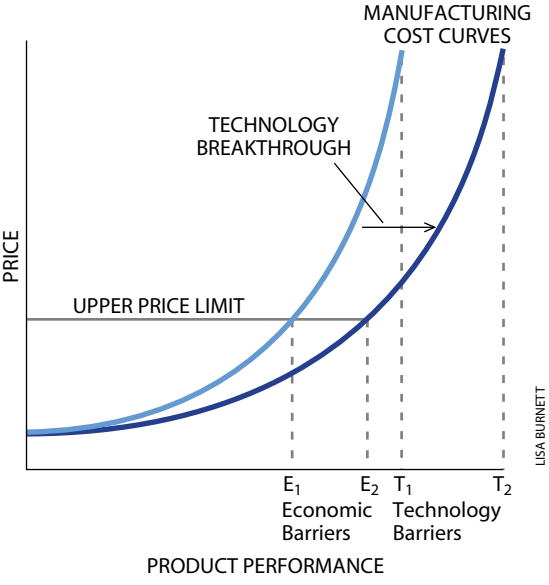
Innovative solutions overcame these limitations. Planarizing methods were developed to ensure optically flat surfaces. Fine adjustments to the edges of the patterns in the reticle were used to shift the phase of the incoming i-line radiation, permitting crisper edge definitions and therefore smaller features—in effect, circumventing Rayleigh's limit. One of the last adjustments was the simple acceptance of a lower value of the proportionality constant, which is related to the degree of contrast in the image projected onto the wafer during lithography. For i-line operation, manufacturers gritted their teeth and accepted a lower proportionality constant than was previously thought practical. Use of the lower value meant that the margins during fabrication would be lower, requiring tighter controls over processes—lithography, deposition and etching, for example—to keep the number of acceptable chips per wafer (the yield) high. As a result of these innovations, i-line steppers are now routinely used to expose 0.35-micron features.

In this last instance, what was really at issue was the loss in contrast ratio that a company was willing to tolerate. With perfect contrast, the image that is created on the photoresist is sharp. Like so many of the limitations in the industry, contrast ratio was perceived to be a technical barrier, but it was actually a risk decision. Lower contrast ratios did not lower yields, it was found, if there were tighter controls elsewhere in the process.

It has been difficult to predict when—or if—this stream of creative improvements will dry up. Nevertheless, as the stream becomes a trickle, the economic consequences of approaching technical barriers will be felt before the barriers themselves are reached. For example, the costs of achieving higher levels of chip performance rise very rapidly as the limits of a manufacturing technology are approached and then surpassed. Increasing costs may drive prices beyond what buyers are willing to pay, causing the market to stagnate before the actual barriers are encountered.

Eventually, though, as a new manufacturing technology takes hold, the costs of fabricating chips begin to decline. At this point, the industry has jumped from a cost-performance curve associated with the old technology to a new curve for the new process. In effect, the breakthrough from one manufacturing technology to another forces the cost curve to bend downward, pushing technical limits farther out [*see illustration below*]. When this happens, higher levels of performance are obtainable without an increase in cost, prompt-

PRICE VERSUS PERFORMANCE



SOURCE: VLSI Research, Inc.

COST CURVE is associated with each chip-manufacturing system. Technology barriers, T_1 and T_2, are where minute increases in chip performance can be achieved only at a huge cost. Economic barriers are encountered well before the technological ones, however. These occur where the line representing the maximum price customers are willing to pay intersects with the curves (at E_1 and E_2). Technology breakthroughs have the effect of bending the curve downward, to the position of the darker plot. When this happens, performance improves, shifting the barriers to E_2 and T_2.

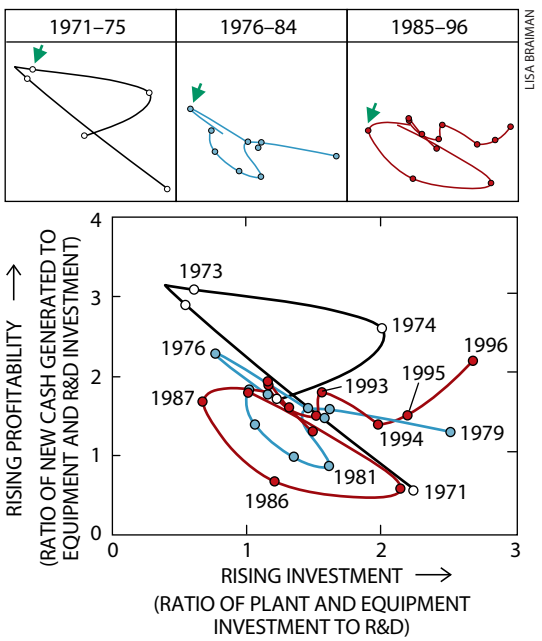
Copyright 1997 Scientific American, Inc.

How Much Bang for the Buck?

For about 60 years, almost all industrial companies have used basically the same model to keep track of financial returns from their investments in equipment, research, marketing and all other categories. Developed just before World War I by Donaldson Brown of Du Pont, the model was brought into the business mainstream by General Motors during its effort to surpass Ford Motor Company as the dominant maker of automobiles.

Since its universal adaptation, this return-on-investment (ROI) model has held up well in industries in which the rates of growth and technological advance are relatively small. To our knowledge, however, the model has never been shown to work well in a sector such as the semiconductor industry, in which many key rates of change—of product performance and the cost of manufacturing equipment, to name just two—are in fact nonlinear. From an economic viewpoint, it is this nonlinearity that makes the semiconductor industry essentially unlike all other large industries and therefore renders unsuitable all of the business models used in other industries to track investment and profitability.

In the semiconductor industry, relatively large infusions of capital must be periodically bestowed on equipment and research, with each infusion exponentially larger than the one before. Moreover, as is true for any company, investments in research, new equipment and the like must eventually generate a healthy profit. At present, however, semiconductor companies have no way of determining precisely the proportion of their financial returns that comes from their technology investments. This inability poses a serious problem for the semiconductor industry. So for several years we have been working on methods of characterizing the industry that take into account these nonlinear elements, with an eye toward modifying the ROI model.



SOURCE: VLSI Research, Inc.

PHASE CHART shows the relation between Intel's profits and investments in technology throughout the company's history. Plotted points trace loops that each correspond to roughly a six-year cycle. (The cycles are shown in different colors.) During each of them, Intel roams from a period of unprofitable operations caused by heavy investment to an interval of very good cash generation stemming from much lighter investment. Green arrows indicate the year in each cycle when Intel made the most money and spent lightly on equipment.

In the conventional model, additional capital investments are made only when gaps occur between a manufacturer's actual and anticipated capacity (the latter is the capacity a company thinks it will need to meet demand in the near future). Such gaps usually result from the aging of equipment and the departure of experienced personnel. In industries such as semiconductors, on the other hand, not only must increases in capacity be constantly anticipated, but also great advances in the manufacturing technology itself must be foreseen and planned for.

To account for this technology-drag effect, we began by considering the ratio of cash generated during any given year to investments made in new technology the year before. New technology, in this context, consists of both new manufacturing equipment and research and development. Cash generated during the year is the gross profit generated by operations, including money earmarked for reinvestment in R&D. (For tax reasons, the standard practice in the industry is not to include R&D funds in this category but rather to treat them as an operating expense.)

What this ratio indicates are incremental profits per incremental investment, one year removed. It shows, in effect, how high a company is keeping its head above water, with respect to profits, thanks to its investment in ever more costly technology. ROI, in contrast, measures the incremental profits over a year coming from all investments, rather than just those of the previous year.

So far we have merely lumped new manufacturing equipment and R&D together as new technology. But the effect of technology drag becomes more striking when the two categories are separated, and the ebb and flow between them is elucidated. One way of doing this is to compute the ratio of these two investments year by year and then plot it against our old standby: the ratio of cash generated during a given year to investments made in new technology during the previous year. Results for Intel over most of its history are plotted in the chart at the left.

Several interesting aspects of Intel's financial history emerge in this diagram, called a phase chart. Connecting the plotted points traces loops that each correspond to roughly a six-year cycle, during which Intel roams from a period of unprofitable operations caused by heavy capital investment to an interval of very good cash generation stemming from much lighter capital investment. From the chart, it is clear that Intel is now entering another period of heavy capital investment. Other semiconductor (and comparable high-technology) companies go through similar cycles. Of course, the timing of the periods of profitability and heavy investment varies from company to company.

Each loop is lower than the one that preceded it. This insight is perhaps the most significant that the illustration has to offer, because it means that Intel's profits, relative to the capital expenditures generating them, are declining with each successive cycle. Because it shows the full cycle between investment in technology and its payoff, this phase chart is a powerful tool for observing and managing the investment cycles peculiar to this unique, dynamic industry. *—G.D.H. and J.D.H.*

Copyright 1997 Scientific American, Inc.

ing buyers to replace older equipment. This is important in the electronics industry, because products seldom wear out before becoming obsolete.

The principles outlined so far apply to all kinds of chips, but memory is the highest-volume business and is in some ways the most significant. From about $550,000 25 years ago, the price of a megabyte of semiconductor memory has declined to just $4 today. But over the same period, the cost of building a factory to manufacture such memory chips has risen from less than $4 million to a little more than $1.3 billion, putting the business beyond the reach of all but a few very large firms. Such skyrocketing costs, propelled mainly by the expense of having to achieve ever more imposing technical breakthroughs, have once again focused attention on limits in the semiconductor industry.

Breakthroughs Needed

The semiconductor industry is not likely to come screeching to a halt anytime soon. But the barriers now being approached are so high that getting beyond them will probably cause more far-reaching changes than did previous cycles of this kind. To understand why requires outlining some details about the obstacles themselves.

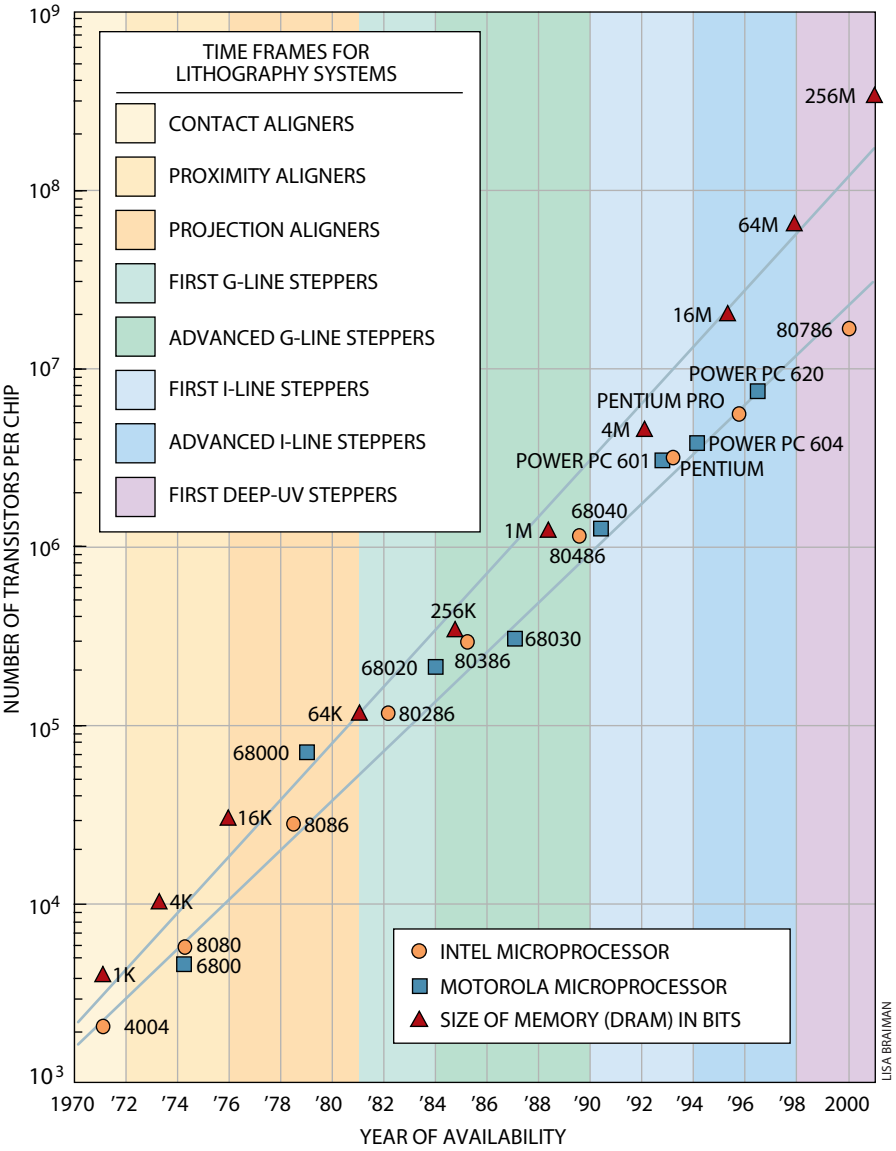
Most have to do with the thin-film structures composing the integrated circuit or with the light sources needed to make the extremely thin conducting lines or with the line widths themselves. Two examples concern the dielectric constant of the insulating thin films. The dielectric constant is an electrical property that indicates, among other things, the ability of an insulating film to keep signals from straying between the narrowly spaced conducting lines on a chip. Yet as more transistors are integrated onto a chip, these films are packed closer together, and cross-talk between signal lines becomes worse.

One possible solution is to reduce the value of the dielectric constant, making the insulator more impermeable to cross-talk. This, in turn, initiates a twofold search, one for new materials with lower dielectric constants, the other for new film structures that can reduce further the overall dielectric constant. Some engineers are even looking for ways to riddle the insulating film with small voids, to take advantage of the very low dielectric constant of air or a vacuum.

Elsewhere on the chip, materials with the opposite property—a high dielectric constant—are needed. Most integrated circuits require capacitors. In a semiconductor dynamic random-access memory (DRAM), for instance, each bit is actually stored in a capacitor, a device capable of retaining an electrical charge. (A charged capacitor represents binary 1, and an uncharged capacitor is 0.) Typically, the amount of capacitance that is available on a chip is never quite enough. Capacitance is proportional to the dielectric constant, so DRAMs and similar chips need materials of a high dielectric constant.

The quest for more advanced light sources for lithography is also daunting. Finer resolution demands shorter wavelengths. But the most popular mercury light sources in use today emit very little energy at wavelengths shorter than the i line's 365 nanometers. Excimer lasers are useful down to about 193 nanometers but generate little energy below that wavelength. In recent years, excimer-laser lithography has been used to fabricate some special-purpose, high-performance chips in small batches.

For still shorter wavelengths, x-ray sources are considered the last resort. Nevertheless, 20 years of research on x-ray lithography has produced only mod-



SOURCES: VLSI Research, Inc.; Integrated Circuit Engineering Corporation

TRANSISTOR DENSITIES on integrated circuits have increased at an exponential rate, as shown on this logarithmic plot. The rate has been sustained by a succession of lithography systems, which are used in chipmaking to project patterns onto wafers. Higher densities have been achieved in memory chips because of their more regular and straightforward design.

Copyright 1997 Scientific American, Inc.

est results. No commercially available chips have as yet been made with x-rays.

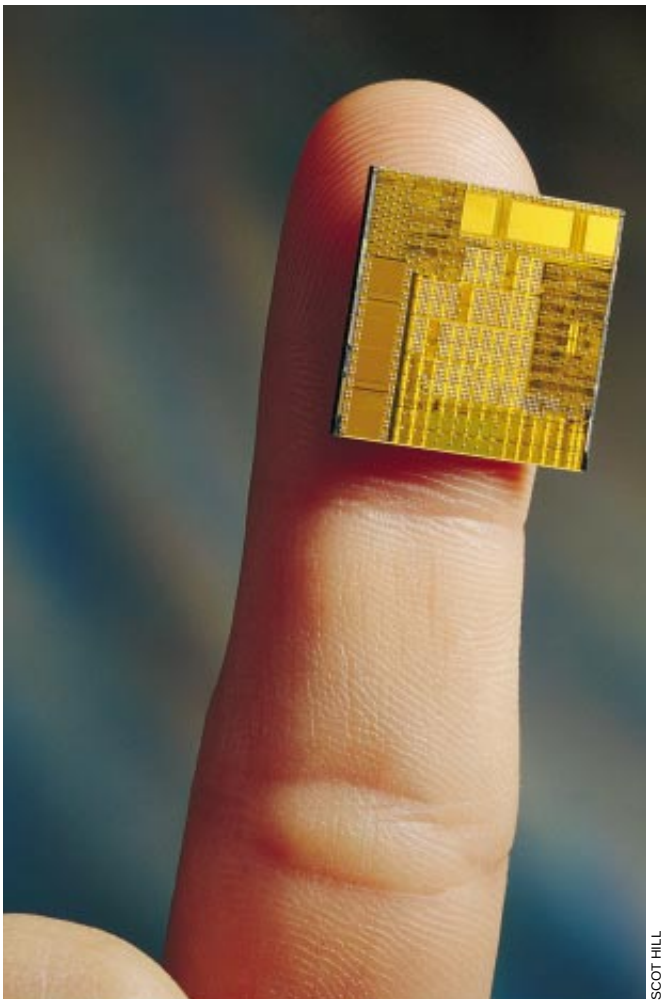
Billion-Dollar Factories

Economic barriers also rise with increasing technical hurdles and usually make themselves evident in the form of higher costs for equipment, particularly for lithography. Advances in lithography equipment are especially important because they determine the smallest features that can be created on chips. Although the size of these smallest possible features has shrunk at roughly 14 percent annually since the earliest days of the industry, the price of lithography equipment has risen at 28 percent a year.

In the early days, each new generation of lithography equipment cost about 10 times as much as the previous one did. Since then, the intergenerational development of stepping aligners has reduced these steep price increases to a mere doubling of price with each new significant development. The price of other kinds of semiconductor-fabrication equipment has gone up in a similar fashion.

Such increases have boosted the overall costs of building semiconductor plants at about half the rate of Moore's Law, doubling every three years. Intel is spending $1.3 billion on its new factory in Hillsboro, Ore., and the same amount on another one in Chandler, Ariz. Advanced Micro Devices (AMD) and Samsung are each building plants that will cost $1.5 billion to finish; Motorola and LG Semicon of Korea have plans for a factory that will cost over $2 billion. Smaller factories can be built for less, but they tend to be less efficient.

That factories now cost so much is one piece of widely cited evidence that formidable technical barriers are close. But the fear that the barriers might be insurmountable, bringing the industry to a halt, seems to us to be unfounded. Rather the prices of semiconductors may increase, and the rate of change in the industry may slow.



SCOT HILL

INTEGRATED CIRCUIT, or die, for Motorola's Power PC 620 microprocessor has nearly seven million transistors. It was designed mainly for use in computer workstations and file servers.

Such an occurrence would not be entirely without precedent. The cost per bit of memories rose by 279 percent between 1985 and 1988 without dire consequences. In fact, 1988 was one of the semiconductor industry's best years. When the cost per bit begins to rise permanently, the most likely result will be an industrial phase change that significantly alters business models.

Up, Up and Away

Virtually every industry more than a few decades old has had to endure such phase changes. Although the semiconductor industry is obviously unique, it is still subject to the principles of economics and of supply and demand. Therefore, the history of older, technical industries, such as aviation, railroads and automobiles, would seem to have episodes that could act as pointers about what to expect.

Like the semiconductor industry, aviation had a fast start. In less than 40 years the industry went from the Wright brothers' biplane to the Pan Am Clipper, the Flying Fortress and the Superfortress. Also like the semiconductor industry, aviation initially served mainly military markets before moving on to nonmilitary ones (mail and passenger transport). The aviation industry sustained growth by lowering the costs per passenger-mile traveled, while also reducing transit times. The dual missions are comparable to the semiconductor industry's steadfast efforts to increase the density of transistors on chips and boost performance, while lowering chip costs.

For several decades, aviation grew by focusing its research and development on increasing passenger capacity and airspeed. Eventually, the trends peaked with Boeing's 747 as a benchmark for capacity and the Concorde as one for speed. Although the 747 was a successful aircraft, filling its many seats was often difficult on all but the longest routes. The Concorde, on the other hand, was an economic failure because noise pollution limited its use. Both represented high-water marks, in the sense that technology could not realistically provide more capacity or speed. Nevertheless, aviation did not go into a tailspin. It entered a second phase in which a greater diversity of smaller airplanes were designed and built for more specific markets. The focus of research and development shifted from speed and size to more efficient and quieter operations and more passenger comfort.

In railroads, the trends were similar. From the 19th century until well into the 1970s, the pulling power of locomotives was continually increased in order to lower the costs of moving freight. Locomotives were a significant capital expense, but gains in pulling power occurred more rapidly than increases in cost did. Eventually, however, locomotive development costs became so great that suppliers and users teamed up. The Union Pacific Railroad, the largest rail-

Copyright 1997 Scientific American, Inc.

road of its time, joined with General Motors's Electro-Motive Division to create the EMD DD-40, a monster that turned out to be too big and inflexible for any purpose other than hauling freight clear across the U.S. Its failure led the railroad industry back to the use of smaller engines that could be operated separately for small loads but hitched together for big ones.

Today the semiconductor industry finds itself in a position not unlike that of the railroad companies just before the EMD DD-40. The costs of developing the factories for future generations of memory chips are so high that companies have begun banding together into different groups, each to attack in its own way the enormous problems posed by fabricating these extremely dense chips economically.

Big Plants, Little Variety

From automobile manufacturing, too, come important lessons. In the 1920s Henry Ford built increasingly more efficient factories, culminating with the vast Rouge plant, which first built Model A cars in 1928. Starting with iron ore, the facility manufactured almost all of the parts needed for automobiles. But by that time the automobile industry had already changed, and Ford's efforts to drive down manufacturing costs by building larger and more efficient plants came at the price of product variety. The old joke about Ford was that the buyer could have a car in any color he or she wanted, as long as it was black.

The semiconductor industry could flourish as it encounters technical barriers.

Trends in automobile manufacturing shifted to permit more conveniences, features and models. As the industry matured, Alfred E. Sloan of General Motors recognized that efficiency was no longer increasing with factory size and that big factories were good mainly for building large numbers of the same product. He therefore split the company into divisions with clearly defined markets and dedicated factories to support them. Customers preferred the resulting wider variation in designs, and General Motors was soon gaining market share at the expense of Ford.

Similar scenarios are being played out in chips. Intel split its 486 microprocessor offerings into more than 30 variations. During the early 1980s, in contrast, the company offered just three versions of its 8086 microprocessor and only two versions of its 8088. Dynamic memory chips are being similarly diversified. Toshiba, for example, currently has more than 15 times as many four-megabit DRAM configurations as it had 64-kilobit ones in 1984. The common theme in all these industries, from railroads to semiconductors, is that their initial phase was dominated by efforts to improve performance and to lower cost. In the three transportation industries, which are considerably more mature, a second phase was characterized by product refinement and diversity—similar to what is now starting to happen in chipmaking. Companies are shifting their use of technology from lowering manufacturing costs to enhancing product lines. The important point is that all these industries continued to thrive in spite of higher manufacturing costs.

It may not be long before the semiconductor industry plateaus. The pace of transistor integration will decline, and manufacturing costs will begin to soar. But as the histories of the aviation, railroad and automobile industries suggest, the semiconductor industry could flourish as it encounters unprecedented and even largely impassable economic and technical barriers. In a more mature industry, growth will almost certainly come from refined products in more diversified lines.

Information storage, and those societal functions dependent on it, will keep moving forward. In fact, slowing the rate of progress in semiconductors could have unexpected advantages, such as giving computer architectures and software time to begin assimilating the great leaps in chip performance. Even in the semiconductor industry, maturity can be a splendid asset. SA

The Authors

G. DAN HUTCHESON and JERRY D. HUTCHESON have devoted their careers to advancing semiconductor manufacturing. In 1976 Jerry founded VLSI Research, Inc., as a technical consulting firm. Dan, his son, joined in 1979. They have focused on analyzing how the interplay of technology and economics affects the business of semiconductor manufacture. Before founding VLSI Research, Jerry, who entered the industry in 1959 as a device physicist, held various positions in research, manufacturing and marketing at RCA, Motorola, Signetics, Fairchild and Teledyne Semiconductor. Dan started his career at VLSI Research as an economist, building several simulation models of the manufacturing process. In 1981 he developed the first cost-based model to guide semiconductor companies in their choice of manufacturing equipment. Dan serves on the University of California's Berkeley Extension Advisory Council and is a member of the Semiconductor Industry Association's Technology Roadmap Council.

Further Reading

IS SEMICONDUCTOR MANUFACTURING EQUIPMENT STILL AFFORDABLE? Jerry D. Hutcheson and G. Dan Hutcheson in *Proceedings of the 1993 International Symposium on Semiconductor Manufacturing*. Institute of Electrical and Electronics Engineers, September 1993.

SIA 1994 NATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS. Semiconductor Industry Association, 1994.

LITHOGRAPHY AND THE FUTURE OF MOORE'S LAW. Gordon E. Moore in *SPIE Proceedings on Electron-Beam, X-Ray, EUV, and Ion Beam Lithographies for Manufacturing*, Vol. 2437; February 1995.

AFFORDABILITY CONCERNS IN ADVANCED SEMICONDUCTOR MANUFACTURING: THE NATURE OF INDUSTRIAL LINKAGE. Donald A. Hicks and Steven Brown in *Proceedings of the 1995 International Symposium on Semiconductor Manufacturing*. Institute of Electrical and Electronics Engineers, September 1995.

SOLID STATE. Linda Geppert in *IEEE Spectrum*, Vol. 34, No. 1, pages 55–59; January 1997.

Copyright 1997 Scientific American, Inc.

Gigabit chips are now in the laboratory.
But the critical technology needed for manufacturing
smaller circuits confronts diminishing returns

Toward "Point One"

by Gary Stix, *staff writer*

Around the turn of the decade, semiconductor makers will begin selling random-access memory chips capable of storing a billion bits of data. These "gigabit" chips, which have already been fabricated in the laboratory, will be made up of transistors with an electrical channel, or gate, measuring as small as 0.18 micron in length. "Point one," as the leading decimal is known by the engineering cognoscenti, is small by any standard: it is about a thousandth the width of a human hair.

The ability of the semiconductor industry to achieve this milestone will depend on continuing advances in the greatest mass-production technique of all time. "An electronic memory circuit will have gone from $10 in the 1950s down to a hundred thousandth of a cent a few years from now," says Alec N. Broers, head of the engineering department at the University of Cambridge. "It's gone further than any technology in history."

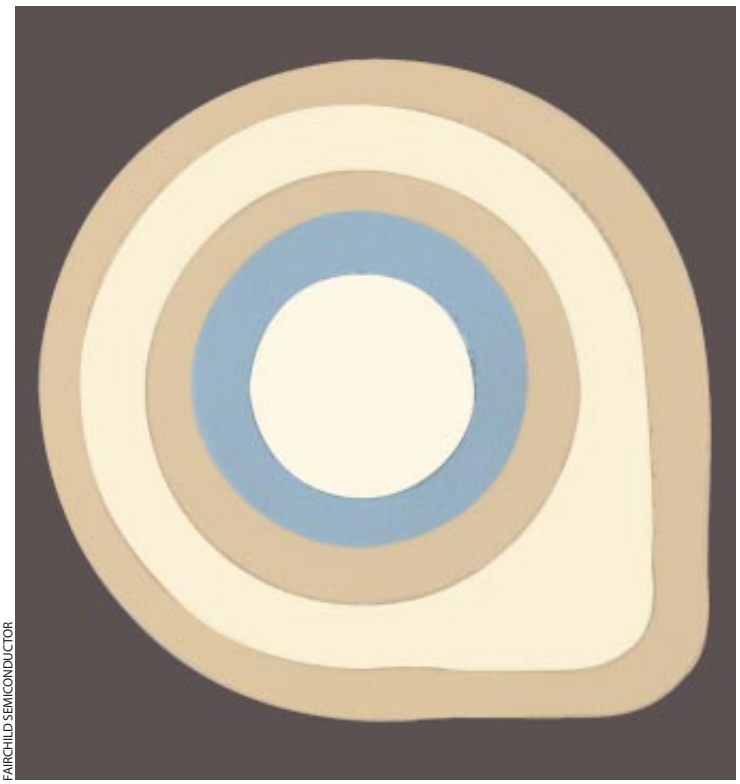
Yet by all accounts, it has become more challenging to make chips faster and smarter by shrinking the size of transistors squeezed onto a few square centimeters of silicon. "We've never had the physics confronted so dramatically as now," says Richard R. Freeman, a lithography expert at Lawrence Livermore National Laboratory.

The physicists, chemists, engineers and materials scientists who study the lithographic-imaging process at the heart of chip manufacturing are having trouble deciding how to advance the technology. The smallest structural dimensions in the most advanced generation of chip technology now being introduced into the marketplace already measure 0.25 micron—and they are getting smaller.

At dimensions near 0.1 micron, the photographic process for developing a circuit image on the surface of a chip starts to falter. Circuit patterns projected through giant $2-million lenses blur easily on the chip's surface. The ultraviolet light focused on the chip gets absorbed before it can print images of submicron transistors, capacitors and wires.

Photolithography has progressed beyond the most optimistic predictions. But if the light finally fades, lithographers may grudgingly have to consider a technology on which numerous corporate and university research careers have turned during a span of more than two decades. Lithography using the short, nanometer wavelengths of x-rays may be the only means of fashioning circuits with billions of transistors. Several years ago a small group of U.S. companies, including IBM and Motorola, launched an effort to share the development costs needed to bring x-ray lithography into the factory.

Champions of x-rays for making chips have not won many converts, even among lithographers within their own companies. In fact, the high costs and technical uncertainties of x-rays have prompted U.S. industry groups to contemplate spending $300 million to $500 million in an effort to bring more conventional optical lithography to the point where it might be used in manufacturing a gigabit memory chip with a billion transistors or a microprocessor that cycles at billions of times per second. "I won't quote anybody, but I was in a meeting where people said that when we get out of optics we're out of the business," says Karen H. Brown, director of lithogra-



FAIRCHILD SEMICONDUCTOR

SMALL GETS SMALLER as witnessed in a then and now comparison of two transistors. The first planar transistor, vintage 1959, measured 764 microns in diameter and could be seen with the naked eye (*left, viewed from above*). One contemporary

Copyright 1998 Scientific American, Inc.

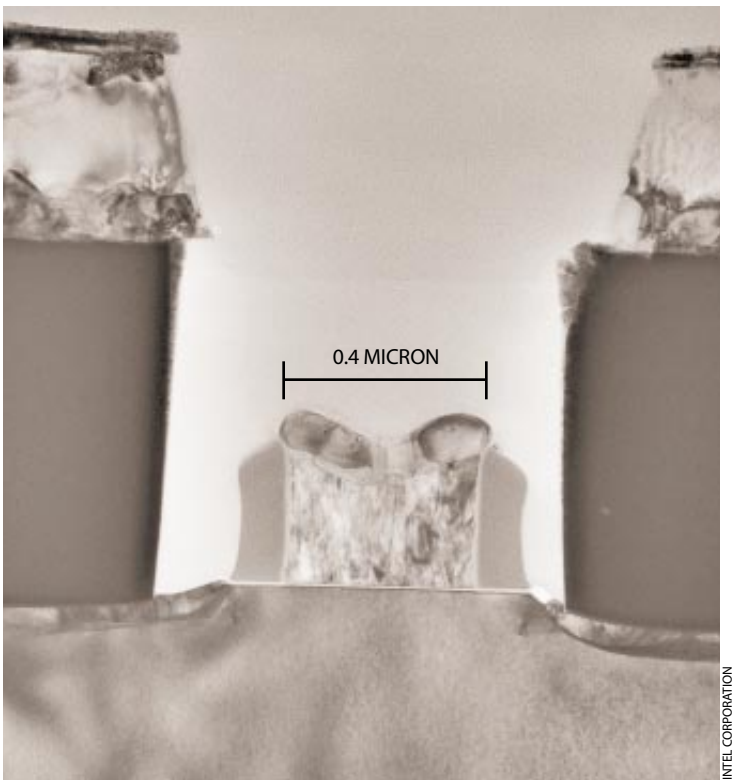
phy for Sematech, the U.S. semiconductor industry's research and development consortium.

Strains between these factions have sometimes led to a tug-of-war to obtain scarce government funding. "Money spent on x-ray lithography has come at the expense of advanced optical lithography," says John R. Carruthers, director of components research at Intel. If some means can be found to let optical technology prevail, the huge x-ray lithography development effort may be written off without its ever having produced a transistor sold on the commercial market.

Tensions between advocates of the two technologies—and inherent doubts about the myriad other technical problems to make circuits with such small dimensions—mean that the relentless three-year cycles for introducing a new generation of memory circuits and microprocessors may start to slow. Worldwide this trend could have a dramatic effect on a $150-billion semiconductor industry that is projected to double in revenues by 2000. The industry may have to find ways of achieving productivity gains beyond making tinier components. Otherwise, the steep decrease in costs for a unit of memory or logic in each successive chip generation could disappear.

From the standpoint of basic physics, the dominant type of chip, the metal oxide semiconductor, might continue to operate down to dimensions of 0.03 micron, about a tenth the size of the most advanced circuitry being manufactured in the factory. Below that scale it may be difficult to turn off the tiny switches called transistors. They would act less like switches than leaky faucets: an electron may move uncontrollably from one side of a transistor to another.

But manufacturing difficulties could cause the technology to expire before then. When millions of transistors are linked to one another, wires must be stacked like a multitiered freeway—electrical resistance in these small-diameter wires and the distance a signal must travel slow operating speeds. Engineers must stand guard against 60-nanometer particles, so-called killer defects that can ruin the memory or brain of a "smart" device. Building chips this small requires very large factories: state-of-the-art manufacturing facilities are headed toward a price tag of $2 billion. Of all these hurdles, however, one of the most daunting is posed by the attempt to wring more out of lithography.



INTEL CORPORATION

transistor, shown in profile (*right*), measures about two microns across and has elements as small as 0.4 micron. Still other transistors in a newer generation of commercial chips incorporate 0.25-micron features.

Photolithography is a hand-me-down from a printing process invented in the 18th century by a German map inspector. It channels light through a mask: a plate of quartz covered with chromium lines that trace a circuit pattern. The light subsequently moves through one of the world's most sophisticated optical devices, a series of 30 or more lens elements that can retail for almost $2 million. These demagnifying lenses reduce the image to one quarter or one fifth its original size and project it onto a few square centimeters of a wafer, a silicon disk roughly 200 millimeters across. The light exposes a micron-thick photosensitive polymer—a photoresist—that is spread over the surface of the wafer. The table on which the wafer sits then "steps" to position another area below the beam. (For that reason, the lithography equipment is called a step-and-repeat tool or, simply, a stepper.)

In the next stage of processing, developing chemicals wash away either the light-exposed or the unexposed parts of the photoresist (depending on the needs of the chipmaker). The circuit pattern on the resist gets transferred to the surface of the wafer by the action of etching chemicals. Lithography serves multiple purposes. Besides just designating the parts of a transistor, it also delineates where impurities, such as boron or arsenic, should be implanted into a chip to alter the electrical conductivity of circuit elements, a process called doping. Lithography can also define the areas to place metal wires to connect circuit elements. The finished wafer is then cut up into individual chips that are packaged in a ceramic or plastic covering.

Limits of Lithography

The gigabit-chip generation may finally force technologists up against the limits of optical lithography. To make these chips, the industry has targeted lithography that uses a pulsed (excimer) laser that emits light at a wavelength as small as 0.193 micron, in the deep ultraviolet segment of the electromagnetic spectrum. But for wavelengths below 0.2 micron, the photoresists absorb so much of the light that it takes more time to transfer a pattern to a chip. It may be impossible to process economically the many hundreds of wafers an hour produced in a large factory.

The few materials that have been identified for 0.193-micron lens manufacturing perform poorly. The fused silica glass for lenses tends to absorb the illumination and to heat up, which can degrade an image by changing the angle at which the lens refracts the light.

Lithographers confront the formidable task of building structures smaller than the wavelength of light. Wavelengths of 0.248 or 0.193 micron have been targeted to make the 0.180-micron structures for gigabit chips. "Think of it as trying to paint a line that is smaller than the width of the paintbrush," says Steven D. Berger, director of engineering at Integrated Solutions, a stepper manufacturer. "There are ways of

Copyright 1998 Scientific American, Inc.

doing it, but not very many ways of doing it in a controlled manner. The thing with lithography is not doing it once. It's doing it 10^{10} times in one second."

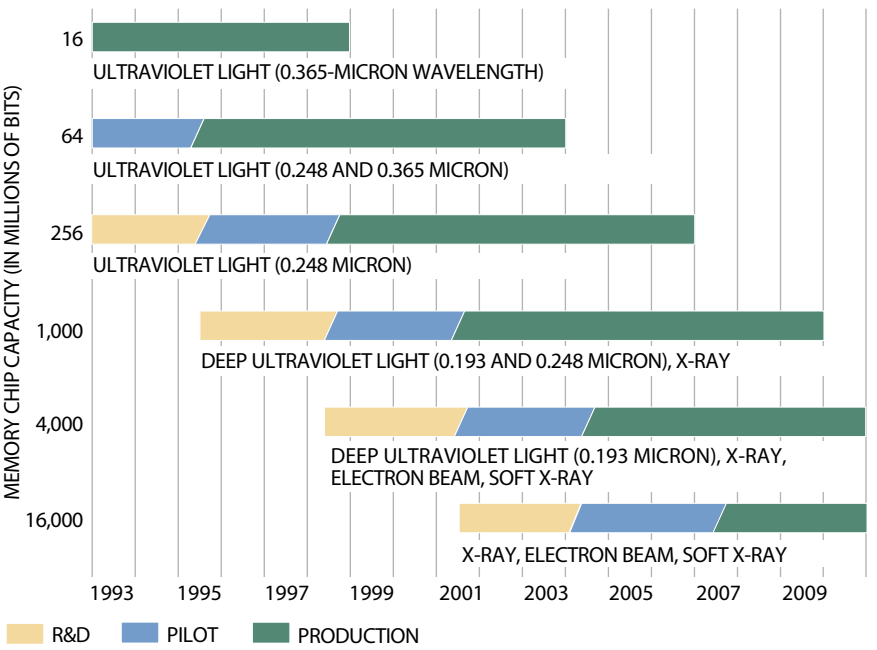
Other problems loom when working at these short wavelengths. Chipmakers must contend with a basic law of optics known by any photographer. A wide lens aperture increases resolution—the strands of a child's hair or tinier chip components stay clearly imaged. At the same time, depth of field decreases. For the amateur photographer, this tradeoff means that the cabin in the background gets a little fuzzy. For the photolithographer, watching over lenses with giant apertures, the focus of the projected image starts to fade at distances well below a micron. It fails to remain sharp down into the jagged Himalayan microtopography of a gigabit chip. As a result, the number of defective chips could skyrocket. "If the yield falls off, you wouldn't have a cost-effective system," says David C. Shaver, head of the solid-state division at the Massachusetts Institute of Technology Lincoln Laboratory.

Light specialists contemplate using what they call tricks, which allow printing of smaller features without reducing depth of field. One scheme employs masks that alter the phase of the light passing through them, which can improve line resolution by 50 to 100 percent. These phase-shift masks are expensive and difficult to make and cannot be used for all the geometric patterns printed on the resist. Intel, however, is contemplating using some of these techniques, beginning in 1999, to make a new microprocessor whose smallest dimensions measure 0.18 micron.

This depth-of-field problem has caused chemists to consider novel approaches for the photoresist. Surface imaging allows only 0.2 micron or so of the top layer of resist to be exposed, instead of the more typical one-micron depth. After the resist polymer is exposed, it is put into a vacuum chamber and comes into contact with a gaseous compound that contains silicon. In one type of resist the nonexposed areas absorb the silicon, which acts as a barrier that protects the underlying layer from a gas of oxygen ions. The gas etches away the photoexposed sections on the chip. It also etches more deeply into the exposed areas of the resist than just the thin layer that was initially exposed. Besides improving resists, chip manufacturers also try to deal with the depth-of-field problem by polishing, or planarizing, the top layer on a chip with a chemical slurry—it is easier to focus on a flat surface.

Researchers at the M.I.T. Lincoln Laboratory have demonstrated a prototype of a lithography system using light with a wavelength of 0.193 micron. But prototypes are not enough. "They make images; they print stuff," says Sematech's Brown. "We have to go past that to make commercial masks and resists you can buy." For that reason, Sematech has banded with industry, academia and government to explore the building of a manufacturing base to make chips with light from the outer reaches of the ultraviolet spectrum.

Research on 0.193-micron lithography has lagged behind schedule for bringing gigabit chips to market in the early part of the next decade. Whether the process can produce millions of chips a year remains unknown. "There's a chance that 0.193 may prove to be impractical," says David A. Markle, vice president of advanced technology for Ultratech Stepper, a California equipment maker. "We need to introduce a new technology that ultimately supplants conventional optical technology, and it's not clear what it will be."



LITHOGRAPHY DEVELOPMENT in the U.S. calls for increasingly smaller wavelengths of light and other forms of electromagnetic energy to produce chips with ever larger memory capacity. The procession of chip generations will require moving to ever smaller wavelengths of ultraviolet light and then perhaps to x-rays or electron beams.

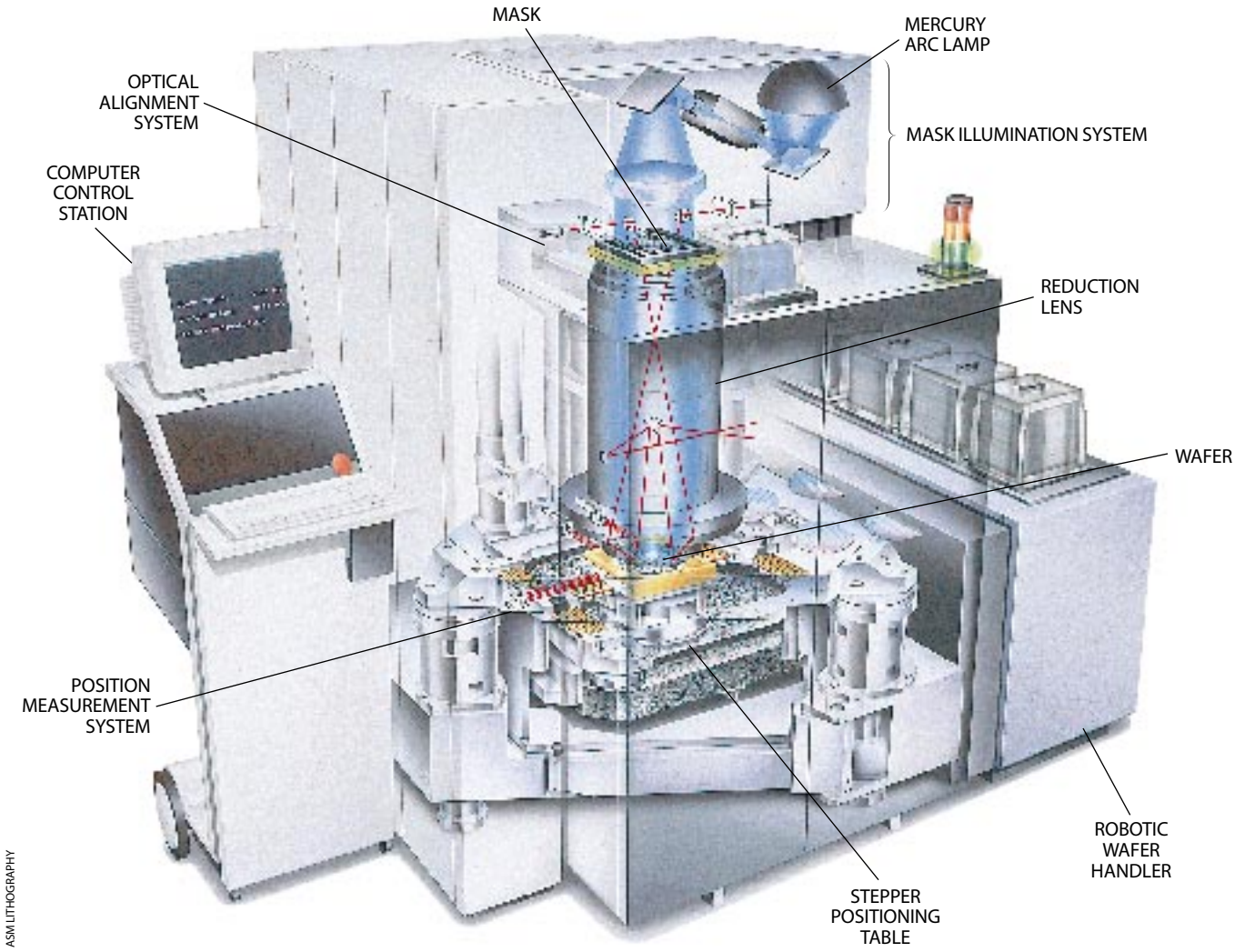
X-ray Visions

A group of researchers who have been closeted in laboratories for decades hope that lithography involving x-rays can succeed if conventional optics fail. X-ray lithography has been nurtured by the Department of Defense's desire for high-performance chips and by IBM, a company that once took it upon itself to develop new semiconductor manufacturing technology. For its part, IBM has acknowledged that it can no longer tread this route alone. It joined with Motorola, Lucent Technologies and Lockheed Martin Federal Systems in setting up a collaboration to move x-ray lithography beyond the laboratory.

In principle, the technology should be a natural choice for drawing finer circuit elements. At roughly one nanometer (a billionth of a meter), x-rays have a wavelength about one four-hundredth that of the light used in today's most advanced commercial systems.

The technology for producing and harnessing x-rays is considerably different from that currently installed on semiconductor fabrication lines—and that is where the debate over the feasibility of x-rays begins. Whereas the radiation for optical lithography can be generated by lasers, the necessary x-rays will most likely emanate from a synchrotron, an energy source usually deployed for physics experiments. IBM owns the only

Copyright 1998 Scientific American, Inc.



ASM LITHOGRAPHY

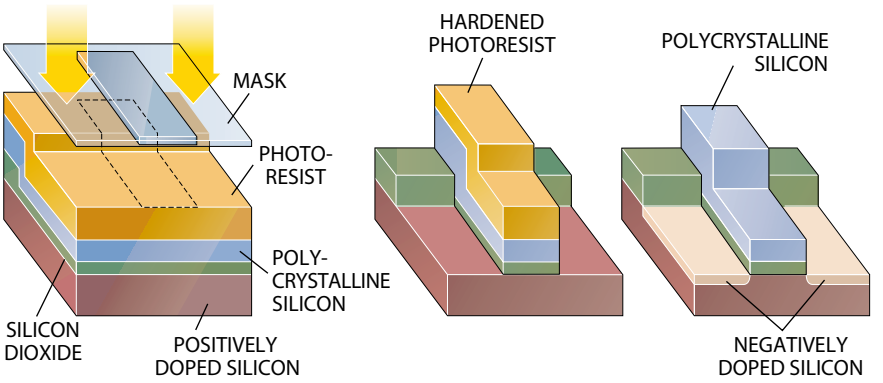
STEPPER, or photolithography machine, imprints circuit patterns on silicon wafers. Ultraviolet light from an arc lamp (or from a laser) passes through a mask bearing the image of the circuit. A sophisticated lens apparatus reduces the image and projects it onto part of a wafer. The table then moves, or "steps," to expose other chips.

commercial synchrotron storage ring in the U.S. It consists of two superconducting magnets whose field confines electrons within a closed orbit. Electrons emit x-rays as they circulate within the storage ring. (Several Japanese companies are also working on x-ray lithography development.)

The price tag of $20 million to $50 million for such x-ray generators should not deter their purchase. Those amounts are 3 percent or less of the cost of the newest semiconductor plants. Moreover, a synchrotron can supply x-rays to 16 steppers simultaneously. For a sizable increase in plant capacity, a $20-million outlay is not unreasonable. A more imposing challenge stems from the need to redesign the entire plant around the x-ray lithographic process. One problem is that if a company wants to make a *small* increase in plant capacity, that, too, costs $20 million.

Another technical obstacle springs from the lack of a commercially feasible way to focus x-rays. Given that x-ray steppers lack the equivalent of lenses (or

PATTERNING A TRANSISTOR involves photolithography in which light is projected through the clear parts of a quartz mask (*left*). The photoresist, a polymer coating, reacts to the light; exposed areas are then removed with a solvent. A plasma of ions etches through the unprotected polycrystalline silicon layer and the silicon dioxide insulating layer (*center*); the rest of the photoresist is removed. The silicon (*pink area*) is implanted with impurities, such as arsenic. Free electrons in this "negatively doped" area conduct current (*right*).



STEVEN STANKIEWICZ

Copyright 1998 Scientific American, Inc.

equivalently demagnifying mirrors), the masks must bear more of the engineering burden: each circuit element outlined by the mask has to be the same small size as the structure to be created on the chip.

The inability to reduce the image also complicates the process of aligning one mask image atop another. Making a gigabit chip requires 20 or so lithographic steps, each with a separate mask. Up to eight of these steps dictate that the x-ray mask alignments be precise to within a few tens of nanometers—a difficult mechanical tolerance to meet.

Nevertheless, IBM has set about tackling the most imposing challenge to the technology's success. It received funding from the Defense Advanced Research Projects Agency (DARPA) to set up a facility to fabricate commercial x-ray masks. Materials that absorb x-rays are hard to find. Gold and tungsten will do the job. But these metals must be laid down in awkward dimensions atop a membrane of silicon through which the radiation is transmitted. A gold circuit feature may have to be 0.4 or 0.5 micron high to absorb the x-rays but only 0.10 micron wide. It begins to look

"Work on x-ray technology was done too early," wrote x-ray pioneer Eberhard Spiller.

"like the New York City skyline," says Larry F. Thompson, vice president of technology development for Integrated Solutions, a stepper manufacturer.

Uncertainties about optical lithography may keep x-rays going. But time may be running out. Motorola and Lockheed Martin Federal Systems have lent a lingering legitimacy to a development effort carried for years by IBM, which had started to lose support from government funding agencies. IBM, after spending a few hundred million dollars on x-rays, looks back with some regret. "In hindsight, our work in x-ray lithography was done much too early," wrote Eberhard Spiller, a pioneer of the technology, in a 1993 issue of the *IBM Journal of Research and Development.*

ALAN D. WILSON/*IBM*

SYNCHROTRON, the only one of its kind in the U.S. designed for commercial manufacturing, is unloaded into IBM's East Fishkill, N.Y., facility on March 29, 1991.

Alternative Technology

Lithographers might consider x-ray technology more favorably if this form of radiation could be demagnified through a lens. A program to devise an x-ray lens is the goal of one of the most advanced research efforts in lithography. This "projection" x-ray system tries to apply to chipmaking, a technology developed, in part, for the Strategic Defense Initiative.

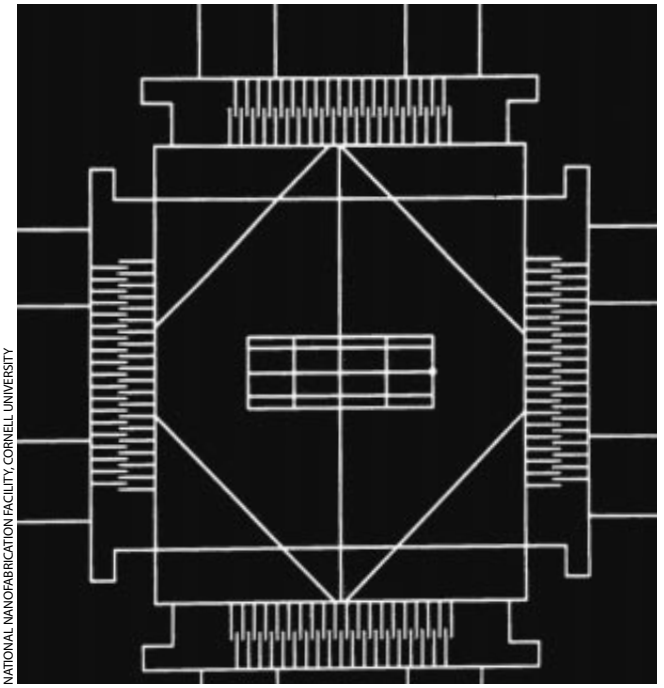
X-ray lithography research is also intended to meet the policy goal of having the government's nuclear weapons and energy laboratories assume a post–cold war role that includes some of the basic research duties once fulfilled by Bell Labs and IBM's laboratories. The laser and measurement expertise of the national labs might be adapted to researching lithography.

Three national laboratories—Sandia, Lawrence Berkeley and Lawrence Livermore—have spearheaded an effort to develop a projection x-ray system for exposing chip elements of 0.1 micron or less. They have worked with Intel and some other companies to provide support, expertise and testing facilities for this program. The consortium's approach is to train a high-powered laser onto a metal target to generate x-rays, which then illuminates a reflective mask. The resulting energy bounces among a series of mirrors that reduce the image to the size of the chip on the wafer. (Nikon and Hitachi are also researching this technology.)

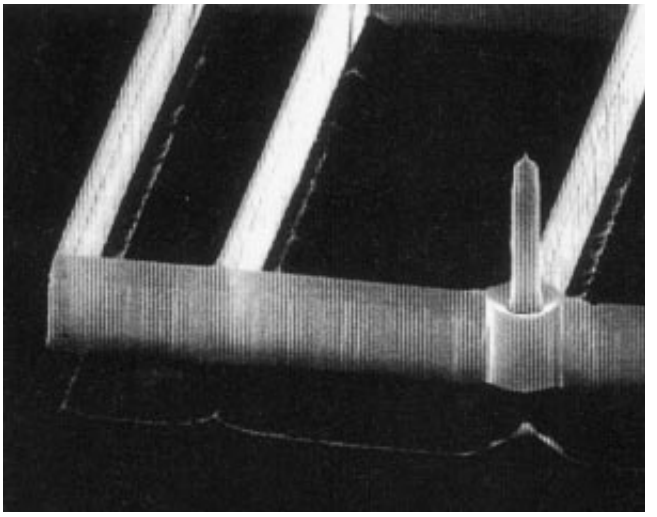
Making microprocessors commercially with this system may be as onerous as tracking and shooting down an incoming ballistic missile. By alternating layers of silicon and molybdenum, researchers have successfully created materials that reflect x-rays instead of absorbing them. They are nonetheless left with the burdensome task of polishing and coating the mirrors to angstrom-level specifications. They must maintain this level of smoothness for each mirror over an area of many square centimeters.

It is a sign of the times that whereas researchers formerly referred to the radiation used in these experiments as "soft" x-rays, they now label it "extreme ultraviolet." The name change reflects the stigma that has come to be attached to x-ray lithography.

Copyright 1998 Scientific American, Inc.



NATIONAL NANOFABRICATION FACILITY, CORNELL UNIVERSITY



FUTURE LITHOGRAPHY is presaged by this 200-micron-wide motor (*left*) from Cornell University that incorporates a 20-nanometer-diameter silicon tip (*close-up shown at right*). The machine's ability to move the tip up, down or sideways could make it a forerunner of lithographic tools that incorporate many tips for patterning a chip surface.

Surprisingly enough, the technology that can make the smallest chip structures is already used every day in manufacturing. Electron-beam lithography employs a focused "pencil" of charged particles to draw lines directly on a photoresist. Indeed, companies sketch circuit patterns onto photolithographic masks with this technique. For 20 years, engineers have dreamed of marshaling it for high-volume lithography. Unfortunately, electron beams are achingly slow: a pencil beam must draw each element of a circuit pattern individually, rather than exposing the entire chip surface in a flash of light. It can take hours to make single chips—too long for mass production, although some high-performance electronics that use nonsilicon materials are made this way.

Since the late 1980s Bell Labs has studied a method that scans a broad electron beam across a chip. As in photolithography, the radiation gets projected through a mask, and the image is reduced with a lens. Lucent Technologies, the parent of Bell Labs, considers this scanning-beam technique to be the most promising for long-term lithography. Still far ahead is a lithographic technique that could promote the embryonic science of nanotechnology. In theory, microscopic tools might fashion the tiniest silicon transistors, those whose smallest dimensions measure only a few tens of nanometers. They might also design new types of electronic devices that store or process information by sensing the position of individual atoms.

Conventional optical lithography can make such tools. It sketches the outlines for hundreds or even thousands of cathodes on silicon. When a voltage is applied to the cathodes, they generate beams of electrons, which can draw circuit lines less than 0.05 micron wide. Noel C. MacDonald, a professor of electrical engineering at Cornell University, has built an array of 1,000 cathodes, providing the makings for an electron-beam machine on a chip.

MacDonald foresees employing the technology in making masks—and perhaps later for actually building small chips. MacDonald, with students Yang Xu and Scott A. Miller, has also demonstrated how a scanning tunneling microscope can be integrated with motors 200 microns wide, which are also fabricated with photolithographic methods. The sharpened tip of the scanning tunneling microscope has been used in research laboratories to push around atoms. The micromotors might let an array of tips—thousands or even a million—pattern a surface rapidly enough for commercial manufacture of circuit lines of but a few nanometers.

Graduate Research

Arrays of cathodes or scanning tunneling microscopes are examples of the most advanced research projects in lithography anywhere. But they are still graduate-level research projects, not equipment that can be bought by leading chip manufacturers.

Perhaps only one or two of these technologies will make it to the factory floor. It simply costs too much to fund any more than that. Expenses for big-ticket lithography may require industry collaborations of competing suppliers, whether their corporate headquarters is in Tokyo or Silicon Valley.

Design ingenuity may overtake the drive to craft diminutive physical attributes. Chips can be made bigger to hold more components. Horizontal layers of transistors could be stacked one atop the other on the same chip to increase memory or logic density. All the while the size of individual chip elements—a transistor or capacitor—may remain the same.

The art and science of lithography may be reaching a mature stage. Motorola or Intel might learn by consulting with retired Boeing executives. In constant dollars, air travel is practically free compared with its cost 35 years ago. But speed of flight has not experienced a similar progression. The expense of building and flying supersonic transports means there are very few of these airplanes carrying passengers today. "We're flying at the same speed that the 707 flew in 1958," says Lawrence Livermore lithography expert Richard Freeman. "I think the same kind of thing could conceivably happen here." In other words, "point one" may be the chip industry's Mach 1. SA

Copyright 1998 Scientific American, Inc.

Tackling Tyranny

by Alan Goldstein

"The only useful thing I was doing was thinking," recalls Jack Kilby in his slow midwestern drawl. A lanky man at six feet, six inches, Kilby leans back in his chair with his hands behind his head and describes the summer that would make him an electronics legend.

The oppressive heat had arrived earlier than usual in Dallas in 1958. Kilby, then a 34-year-old electrical engineer, had been left on his own in the semiconductor laboratory at Texas Instruments, assigned to solve a problem that was stalling the entire electronics industry.

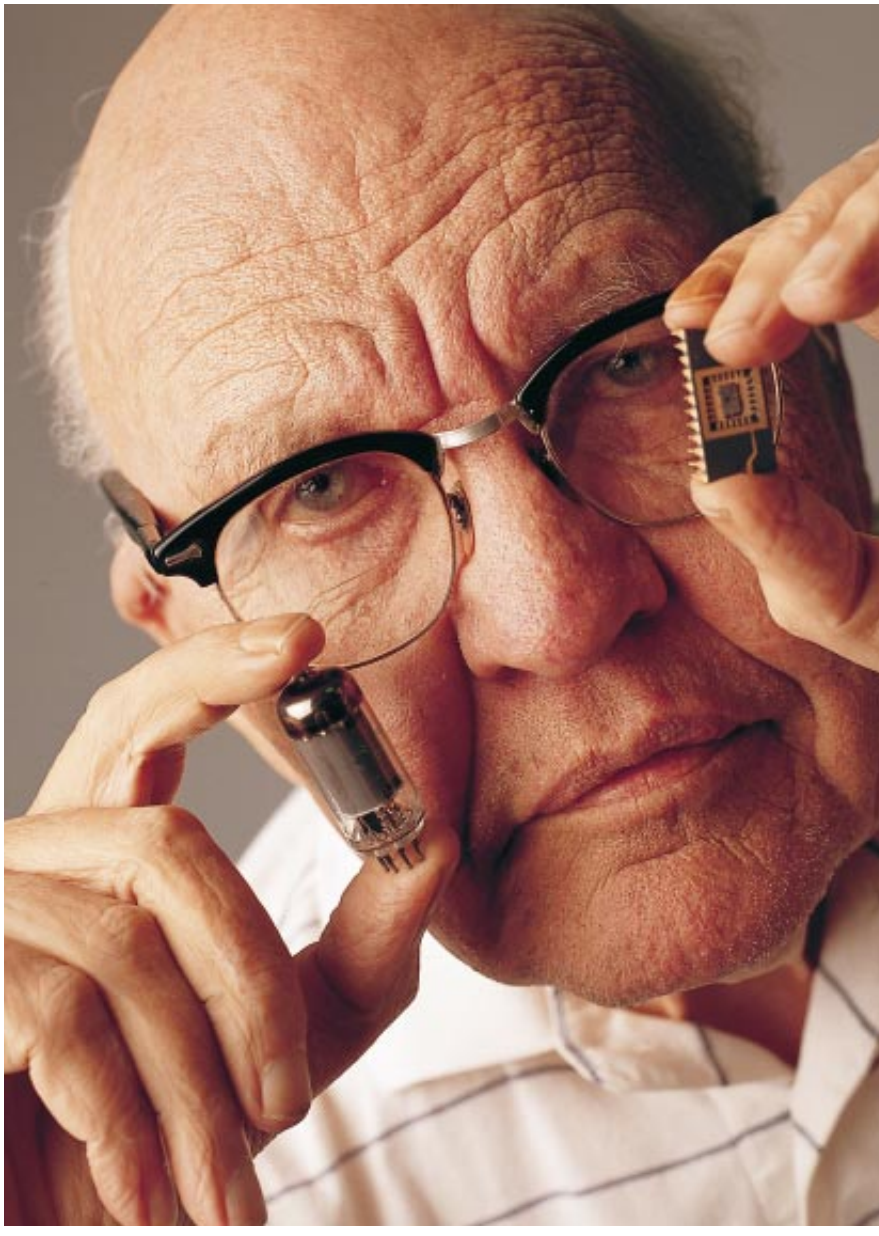
The difficulty was called the "tyranny of numbers"—a succinct name for a complicated problem. The transistor, invented 10 years earlier, made it theoretically possible to build a powerful computer the size of a wristwatch. Unfortunately, such a computer could not be manufactured, because no one had figured out how to link together the vast number of required components. (Today's advanced computer chips, for instance, contain millions of transistors on a surface not even half the size of a postage stamp.)

At the time, the best available method for assembling the electronics—soldering the myriad connections by hand—produced computers that filled entire rooms but had limited capabilities. Computational rates were slowed by the delay of electronic signals traveling through miles of wiring to complete a circuit. Not only was the soldering method expensive, but some electrical connections were inevitably flawed, forcing engineers to design redundant wiring routes that only exacerbated the tyranny of numbers.

Hired by Texas Instruments to tackle this obstacle, Kilby moved his wife and two daughters from Milwaukee (where he had been designing electronic circuits for the company Centralab) to Dallas in May 1958. The adjustment was tough: their apartment was cramped, and they didn't much like the area at first. During that first summer on the job, Kilby was alone in the office not by choice but because of company policy—as a newcomer, Kilby didn't have the seniority to participate in the staff's mass vacation exodus in July.

The idea that ushered in an age of notebook computers, pocket cellular phones and numerous other applications did not come in a single brilliant flash. Instead Kilby recalls that he built on one small advance after another until he came up with the notion of making all parts of the circuit out of a single block of material, such as silicon. That way, he figured, all the components would be connected, or integrated; no matter how complex the circuit, nothing would have to be wired together. The tyranny of numbers would be broken.

Like many great inventions, the monolithic integrated circuit, or microchip, seems simple in retrospect. But it was anything but obvious at the time. When his colleagues came back from vacation, Kilby showed off his notebook sketches. His boss was skeptical but did approve development of a model.



DANNY TURNER

Copyright 1997 Scientific American, Inc.

Jack Kilby,

one of the inventors of the integrated circuit, recalls a hot summer in Dallas that changed the electronics industry forever

On September 12, 1958, a group gathered in Kilby's area of the lab. Kilby remembers being nervous as he hooked the wires from a battery to his small monolithic circuit and from the circuit to an oscilloscope. He took a deep breath and flipped a switch. A green line undulated in a perfect squiggly wave. Everyone smiled. The chip was born.

JACK KILBY helped to usher in the age of microcomputers. Here he holds a vacuum tube (*left*) and a modern integrated circuit (*right*).

All these years later Kilby, now 73 and working as a consultant in a spartan office a few miles from Texas Instruments headquarters in Dallas, has a different perspective. "There was never any real question whether it would work," he says confidently. "I thought if it were successful, it would have an impact on the world of electronics as we knew it then—radio, TV and computers. I had no idea how it would expand the world of electronics."

Indeed, a decade passed before people realized the importance of the chip. And Kilby had to fight to receive credit. A few months after he demonstrated the integrated circuit, the same idea occurred to Robert N. Noyce, an engineer working at a young company in northern California called Fairchild Semiconductor Corporation.

Both men filed for patents. Although Kilby had the paperwork to show he was first, his application contained gaps and ambiguities. The result was a protracted legal fight that ended in a draw. Although the courts leaned toward Noyce, Kilby and Noyce are regarded among engineers as co-inventors of the chip. Noyce went on to help start the chip-manufacturing company Intel Corporation and become fabulously wealthy in the process. (He died of a heart attack in 1990 at the age of 62.)

But the executive suite wasn't for Kilby, a quiet man who didn't care for office politics. After demonstrating the integrated circuit, a series of promotions eventually made him deputy director of the semiconductor research program at Texas Instruments. In 1965 an effort to prove the utility of the microchip to the average person led to the first handheld calculator.

By 1970, though, Kilby was frustrated with life in a big corporation and convinced that creativity required more freedom. So he left to work on his own. None of the patents Kilby has received since leaving corporate life, however, have brought the financial or even the technical success of either the integrated circuit or the handheld calculator. Kilby's independent efforts have included a system for avoiding unwanted telephone calls as well as an electronic checkbook, both of which have now been superseded by other technologies.

Rather than work on new inventions, Kilby now spends most of his time helping clients overcome technological obstacles—a role reminiscent of his assault on the tyranny of numbers. He declines to disclose which companies he collaborates with or to offer much detail about his current work. Kilby lives simply, although a Lexus coupe and a nice home in prosperous north Dallas are evidence of financial comfort.

Does he think that his legacy, the integrated circuit, has improved our lives? "Well, it's certainly changed it, and I suspect for the better," Kilby responds. "Communication between people has become much easier. Cell phones, that sort of thing, are now very common." With characteristic modesty, he quickly adds: "If you don't like the telephone much, and I don't, you may not consider that a plus."

Ever the pragmatist, Kilby is more enamored of solving problems than he is taken with technology. His computer is a Dell 386 model that is several generations old. "This one is still doing everything that I want it to," he assures me. "If I bought a new one, I'd have to learn new software."

Kilby scoffs at some of the future concepts that are kicked around by high-tech executives, like wallet-size PCs or chips incorporated into people's clothing that would exchange information—an electronic business card, in a sense—with a handshake. "It seems boring, not a great way to start a conversation, although it certainly could be done," he says.

Kilby is more concerned that the rapid pace of advances in chip technology will slow. A modern semiconductor factory costs around $1 billion, and the price tag has steadily risen as chips have become more complex. Eventually, he speculates, the cost may exceed the benefits, and something else will have to replace the integrated circuit.

What do we know about what this successor might be? "Almost nothing," he shrugs. That invention will be the challenge for someone else. SA

ALAN GOLDSTEIN is a technology reporter for the Dallas Morning News.

Copyright 1997 Scientific American, Inc.

Silicon may be "God's gift," but to see the light, you'll need gallium arsenide, or indium phosphide, or aluminum indium gallium phosphide, or….

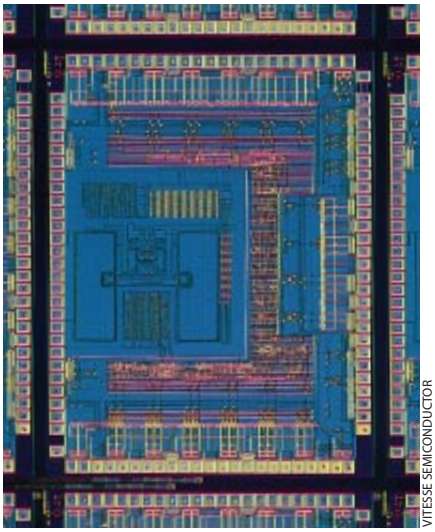
THE SEMICONDUCTING MENAGERIE

by Ivan Amato

Just as the Stone Age, at technology's dawning, was actually an age of stone and bone and hide and wood and whatever else our ancestors found to be useful, this Silicon Age of ours is much more than its moniker suggests. Gallium arsenide, indium phosphide, aluminum indium gallium phosphide and mercury cadmium telluride are but a few of the major players in an ever expanding cast of semiconducting characters that have been finding roles in exotic or ultraefficient lasers and blazingly fast electronic circuits.

Those who think they have never seen or used a nonsilicon semiconductor are wrong. Remember those funky red-light-emitting diodes (LEDs) and watch and calculator displays that first started appearing in the late 1960s? They were based on such materials as gallium arsenide phosphide and gallium phosphide. And LEDs were just the beginning. Practically every compact-disc player contains a two-dollar, mass-produced semiconductor laser, the "stylus" that bounces off the CD's microcode of pits and spaces before shining onto semiconductor photodetectors for eventual conversion into sound. The heart of this laser is an exquisitely thin stripe of gallium arsenide sandwiched between slices of gallium aluminum arsenide, which is more electrically insulating. The infrared light emitted from this sandwich is produced when electrons and positively charged electron deficiencies (called holes) recombine, annihilating one another and releasing photons.

Telephony is another stronghold of nonsilicon semiconductors. In order to transmit effectively at low power, cellular telephones operate at frequencies at the edge of, or beyond, the capabilities of silicon circuitry. Thus, most cellular telephones have gallium arsenide circuitry, within which electrons move faster than they do inside silicon, enabling higher-frequency operation. Hardwired telephone systems, too, rely on nonsilicon semiconductors. Costly high-quality semiconductor lasers made of indium gallium arsenide phosphide, for example, send light-encoded data and voice signals down optical fibers.



VITESSE SEMICONDUCTOR

INTEGRATED CIRCUIT of gallium arsenide contains 25,000 transistors and is used in communications systems.

What most of these applications have in common is that they are all related to the generation and detection of light. "The one thing that silicon cannot do is produce light," laments Harvey Serreze, operations manager of the optoelectronics division of Spire Corporation, a high-tech company based near Boston. Like the majority of those working on nonsilicon semiconductors, Serreze and his colleagues concentrate on the so-called III-V semiconductors, which are made by mixing and matching one or more elements from columns III and V of the periodic table of chemical elements. The usual suspects from column III are aluminum, gallium and indium; column V staples include nitrogen, phosphorus and arsenic. The lure of III-V compound semiconductors—and their even more exotic "II-VI" cousins—is that each has its own set of intrinsic and potentially useful electronic or optical traits.

Underlying these properties is the material's so-called band gap. This term refers to a kind of forbidden territory for electrons associated with the atoms of a crystal. Below the band gap is the valence band, a range of lower-energy electron states in which the electrons in the crystal remain closely associated with a specific atom in the crystal. Above the band gap is the conduction band, consisting of higher-energy states in which electrons are no longer associated with individual atoms and can flow relatively freely. Hence, the band gap is the amount of energy, measured in electron volts, that is required to cause electrons to leap from the valence band, over the band gap and into the conduction band.

Once in the conduction band, the electrons can carry current or fall back across the band gap to the valence band. When this latter event occurs, the electron typically falls into an empty "bonding" site, or electron deficiency, vacated

Copyright 1997 Scientific American, Inc.

by another electron. This phenomenon, known as recombining with a hole, causes a photon to be emitted. Exploitation of these recombinations is the basic principle behind almost all semiconductor lasers.

The key feature in all this activity is the band gap. The energy, and therefore the wavelength, of the emitted photons is a function of the band gap: the wider this gap, the greater the energy involved in falling back across it and the shorter the wavelength of the emitted photon. The band gap also determines whether and under what conditions the semiconductor can detect light and under what range of temperatures it will be able to serve as a reliable part of a transistor in, say, a computer chip.

The chemical composition of each semiconductor determines its intrinsic band gap, but researchers have been coming up with ever more clever ways of coercing these materials to do what they otherwise would not. By sprinkling different amounts of nitrogen atoms into gallium phosphide crystals, for example, materials researchers can alter the crystal's band gap so that it emits either red or green light. Such techniques have given rise to most of the materials—and colors—in the huge, worldwide industry for light-emitting diodes. It has also led to an enormous menagerie of one-of-a-kind materials and devices that make for respectable technical publications but rarely have the right stuff for lucrative ventures.

"People have gone through the periodic table and have come up with exotic [semiconducting] compounds, most of which are too tough to make," Spire's Serreze remarks. Among the pitfalls are too many microscopic defects, which put an end to light-emitting recombinations of electrons and holes, and too much strain within the crystals, which shortens the material's useful lifetime. To avoid the common disappointment of finding coveted properties in impractical materials, theoreticians and experimenters have fostered a golden age in a field known as band-gap engineering.

Band-gap engineers grow fabulously complex semiconductor crystals using such sophisticated, state-of-the-art techniques as chemical vapor deposition and molecular-beam epitaxy. Not uncommonly, these crystals consist of hundreds of layers, many no more than several atoms wide and each consisting of a chemical compound that the crystal grower essentially dials in. A key advantage of band-gap engineering is that it allows researchers to get a desired band gap by varying a crystal's layer-by-layer architecture, rather than by concocting some very esoteric compound that will most likely have undesirable properties that render it useless.

Whatever you call this technological age, semiconductors of many flavors will infiltrate ever more of its nooks and crannies.

Because the layer-by-layer structure, not the specific chemical composition, sets the band gap, engineers are free to use relatively mundane and well-behaved materials, such as gallium arsenide, gallium aluminum arsenide and indium phosphide. With quantum-mechanical reasoning and calculations to guide them, these band-gap engineers design crystals whose interiors are essentially programmed to manipulate electrons passing through them.

Even as nonsilicon semiconductors stake their claim to more territory in technology's landscape, silicon rests secure in its kingdom. For the time being, at least, silicon remains the semiconductor of choice for applications such as computer chips, in which only electronic motion really matters and there is no emission of light. Silicon has proved so fantastic for microelectronic applications that solid-state scientists are often moved to speak of the element in unabashedly reverential terms. "Silicon is God's gift to the world," exclaims Wolfgang Choyke, a physicist at the University of Pittsburgh.

Not that researchers haven't tried and even come close to establishing a gallium arsenide beachhead in silicon's microprocessing stronghold. The legendary computer designer Seymour Cray used gallium arsenide chips in his last machine, the Cray-3 supercomputer. The machine never got off the ground commercially, partly because of the intricate and expensive cooling system needed to counter the heat produced by the chips. "That was such a huge undertaking that it could not find customers," explains Ira Deyhimy, vice president of product development at Vitesse Semiconductor Corporation, one of the largest suppliers of gallium arsenide chips.

At least one company did bring a gallium arsenide–based, high-performance computer to market. In the early 1990s Convex Computer Corporation in Richardson, Tex. (now the Convex Division of Hewlett Packard), shipped more than 60 such machines, each of which sold for several million dollars. According to Steve Wallach, who leads the division, Convex's initial tack was to build their machines using several cooler-running, lower-power gallium arsenide chips, which were also made by Vitesse. The problem, however, was that Vitesse could not produce the chips with a high enough yield to meet Convex's needs (nor could anyone else, Wallach adds).

Deyhimy, however, notes that some supercomputer companies are moving to hybrid architectures that combine the computing power of superdense silicon chips with the swifter chip-to-chip data-shuttling capabilities of gallium arsenide. "Our most complex part in production has 1.4 million transistors," he says, pointing out that even if gallium arsenide's role in microprocessing remains minor compared with silicon's, the material will grow impressively in telecommunications, data communications and other burgeoning information industries that depend more on fast switching speeds and high data rates or frequencies than on processing power.

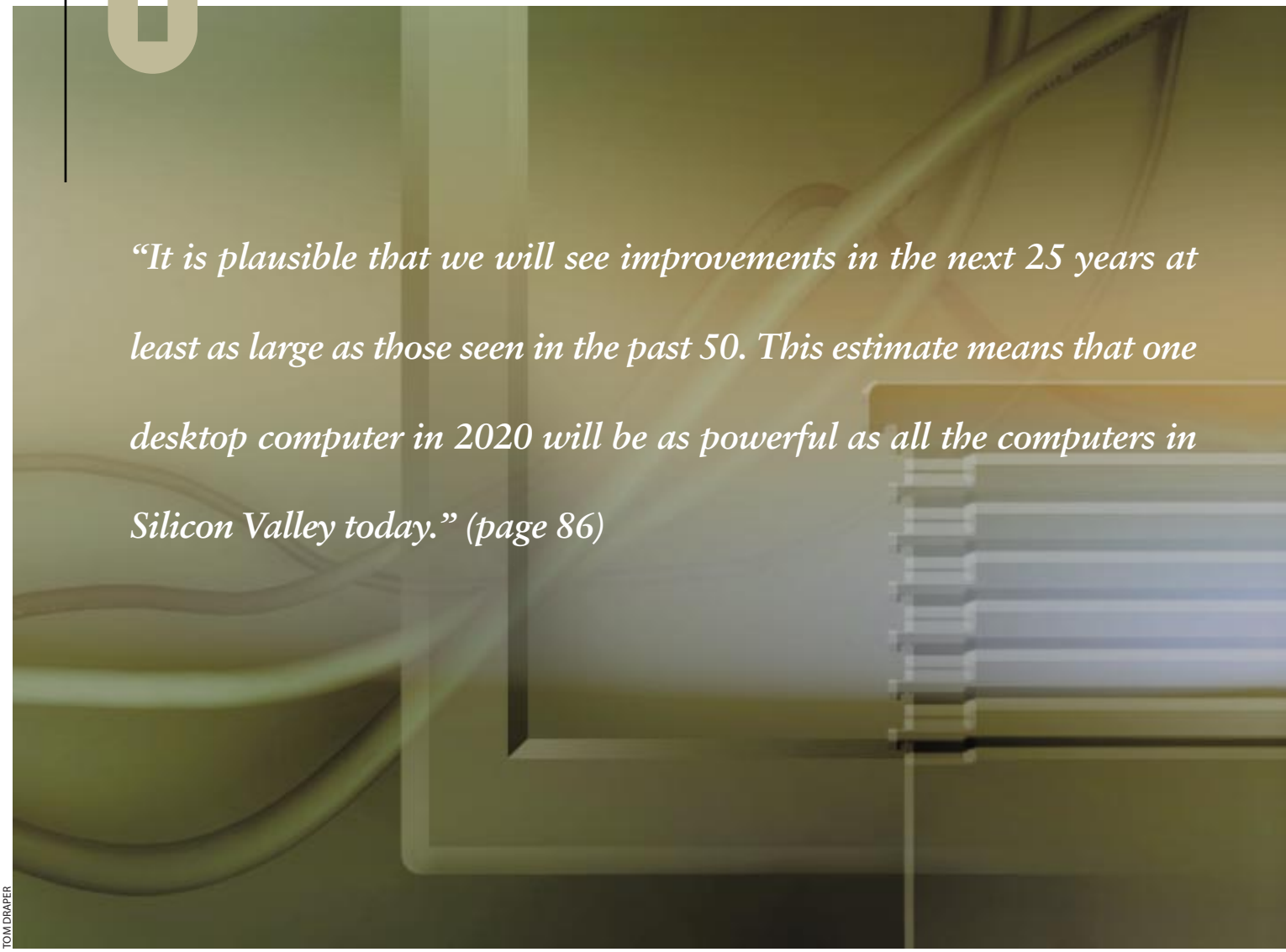
As far as gallium arsenide has come lately in the stratosphere of high-speed electronics for the communications world, it may very well have competition one day from its old nemesis, silicon. The Defense Advanced Research Projects Agency, for one, is funding several research groups that are combining silicon with germanium or with germanium and carbon atoms in various semiconductor alloys and quantum-well structures.

Whatever you call this technological age, semiconductors of many flavors will infiltrate ever more of its nooks and crannies. When it comes to predicting what might be possible, caution has been the hazardous path. Says Serreze, "The naysayers are on the side of the road," whereas the visionaries are speeding by. SA

IVAN AMATO, a freelance writer based in Silver Spring, Md., is the author of Stuff, *published earlier this year by BasicBooks.*

Copyright 1997 Scientific American, Inc.

3 The Revolution Continues



"It is plausible that we will see improvements in the next 25 years at least as large as those seen in the past 50. This estimate means that one desktop computer in 2020 will be as powerful as all the computers in Silicon Valley today." (page 86)

TOM DRAPER

Copyright 1997 Scientific American, Inc.

Every 18 months microprocessors double in speed. Within 25 years, one computer will be as powerful as all those in Silicon Valley today

MICROPROCESSORS IN 2020

by David A. Patterson

Unlike many other technologies that fed our imaginations and then faded away, the computer has transformed our society. There can be little doubt that it will continue to do so for many decades to come. The engine driving this ongoing revolution is the microprocessor, the sliver of silicon that has led to countless inventions, such as portable computers and fax machines, and has added intelligence to modern automobiles and wristwatches. Astonishingly, the performance of microprocessors has improved 25,000 times over since their invention only 27 years ago.

I have been asked to describe the microprocessor of 2020. Such predictions in my opinion tend to overstate the worth of radical, new computing technologies. Hence, I boldly predict that changes will be evolutionary in nature, and not revolutionary. Even so, if the microprocessor continues to improve at its current rate, I cannot help but suggest that 25 years from now these chips will empower revolutionary software to compute wonderful things.

Smaller, Faster and Cheaper

Two inventions sparked the computer revolution. The first was the so-called stored program concept. Every computer system since the late 1940s has adhered to this model, which prescribes a processor for crunching numbers and a memory for storing both data and programs. The advantage in such a system is that, because stored programs can be easily interchanged, the same hardware can perform a variety of tasks. Had computers not been given this flexibility, it is probable that they would not have met with such widespread use. Also, during the late 1940s, researchers invented the transistor. These silicon switches were much smaller than the vacuum tubes used in early circuitry. As such, they enabled workers to create smaller—and faster—electronics.

More than a decade passed before the stored program design and transistors were brought together in the same machine, and it was not until 1971 that the most significant pairing—the Intel 4004—came about. This processor was the first to be built on a single silicon chip, which was no larger than a child's fingernail. Because of its tiny size, it was dubbed a microprocessor. And because it was a single chip, the Intel 4004 was the first processor that could be made inexpensively in bulk.

The method manufacturers have used to mass-produce microprocessors since then is much like baking a pizza: the dough, in this case silicon, starts thin and round. Chemical toppings are added, and the assembly goes into an oven. Heat transforms the toppings into transistors, conductors and insulators. Not surprisingly, the process—which is repeated perhaps 20 times—is considerably more demanding than baking a pizza. One dust particle can damage the tiny transistors. So, too, vibrations from a passing truck can throw the ingredients out of alignment, ruining the end product. But provided that does not happen, the resulting wafer is divided into individual pieces, called chips, and served to customers.

Although this basic recipe is still followed, the production line has made ever cheaper, faster chips over time by churning out larger wafers and smaller transistors. This trend reveals an important principle of microprocessor economics: the more chips made per wafer, the less expensive they are. Larger chips are faster than smaller ones because they can hold more transistors. The recent Intel Pentium II, for example, contains 7.5 million transistors and is much larger than the Intel 4004, which had a mere 2,300 transistors. But larger chips are also more likely to contain flaws. Balancing cost and performance, then, is a significant part of the art of chip design.

Most recently, microprocessors have become more powerful, thanks to a change in the design approach. Following the lead of researchers at universities and laboratories across the U.S., commercial chip designers now take a quantitative approach to computer architecture. Careful experiments precede hardware development, and engineers use sensible metrics to judge their success. Computer companies acted in concert to adopt this design strategy during the 1980s, and as a result, the rate of improvement in microprocessor technology has risen from 35 percent a year only a decade ago to its current high of approximately 55 percent a year, or almost 4 percent each month. Processors are now four times faster than had been predicted in the early 1980s; it is as if our wish were granted, and we now have machines from the year 2002.

Pipelined, Superscalar and Parallel

In addition to progress made on the production line and in silicon technology, microprocessors have benefited from recent gains on the drawing board. These breakthroughs will undoubtedly lead to further advancements in the near future. One key technique is called pipelining. Anyone who has done laundry has intuitively used this tactic. The nonpipelined approach is as follows: place a load of clothes in the washer. When the washer is done, place the load into the dryer. When

Copyright 1997 Scientific American, Inc.

the dryer is finished, fold the clothes. After the clothes are put away, start all over again. If it takes an hour to do one load this way, 20 loads take 20 hours.

The pipelined approach is much quicker. As soon as the first load is in the dryer, the second dirty load goes into the washer, and so on. All the stages operate concurrently. The pipelining paradox is that it takes the same amount of time to clean a single dirty sock by either method. Yet pipelining is faster in that more loads are finished per hour. In fact, assuming that each stage takes the same amount of time, the time saved by pipelining is proportional to the number of stages involved. In our example, pipelined laundry has four stages, so it would be nearly four times faster than nonpipelined laundry. Twenty loads would take roughly five hours.

Similarly, pipelining makes for much faster microprocessors. Chip designers pipeline the instructions, or low-level commands, given to the hardware. The first pipelined microprocessors used a five-stage pipeline. (The number of stages completed each second is given by the so-called clock rate. A personal computer with a 200-megahertz clock then executes 200 million stages per second.) Because the speedup from pipelining equals the number of stages, recent microprocessors have adopted eight or more stage pipelines. One 1997 microprocessor uses this deeper pipeline to achieve a 600-megahertz clock rate. As machines head toward the next century, we can expect pipelines having even more stages and higher clock rates.

Also in the interest of making faster chips, designers have begun to include more hardware to process more tasks at each stage of a pipeline. The buzzword "superscalar" is commonly used to describe this approach. A superscalar laundromat, for example, would use a professional machine that could, say, wash three loads at once. Modern superscalar microprocessors try to perform anywhere from three to six instructions in each stage. Hence, a 250-megahertz, four-way superscalar microprocessor can execute a billion instructions per second. A 21st-century microprocessor may well launch up to dozens of instructions in each stage.

Despite such potential, improvements in processing chips are ineffectual unless they are matched by similar gains in memory chips. Since random-access memory (RAM) on a chip became widely available in the mid-1970s, its capacity has grown fourfold every three years. But memory speed has not increased at anywhere near this rate. The gap between the top speed of processors and the top speed of memories is widening.

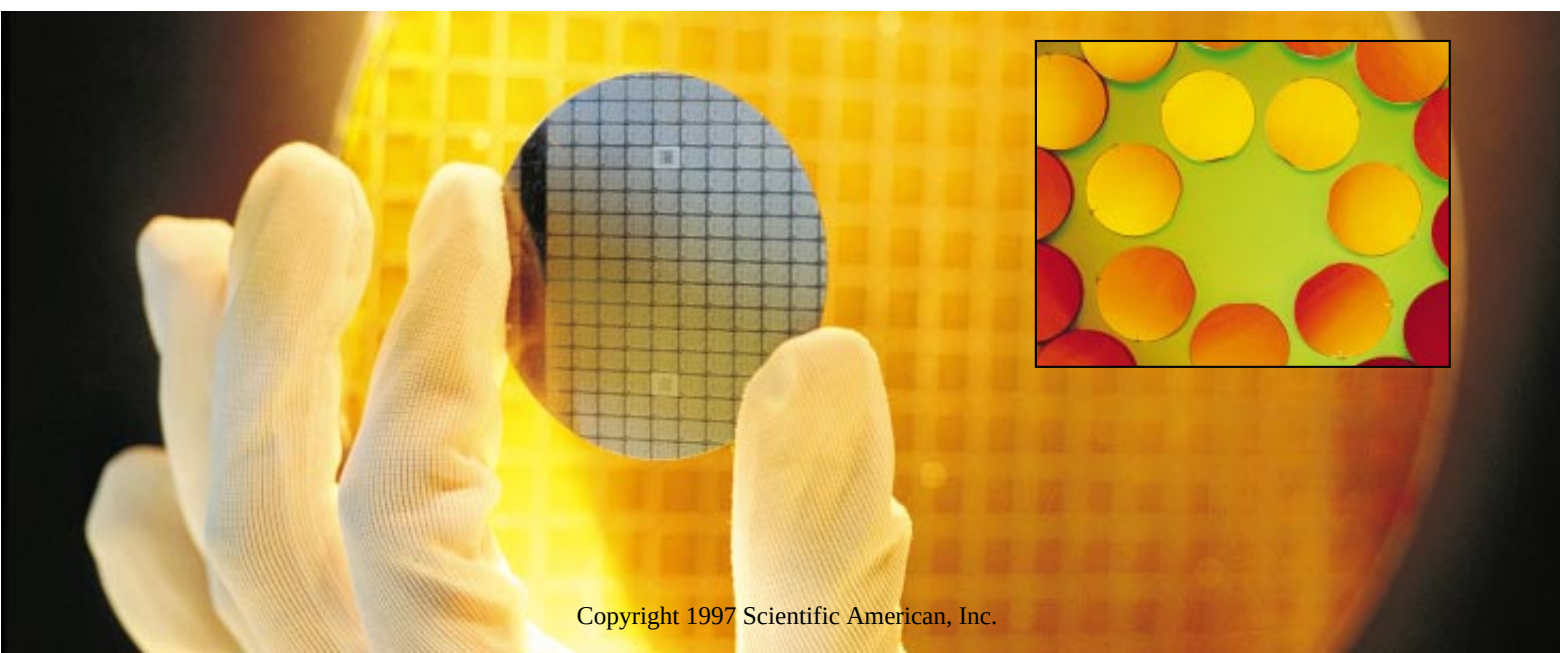
One popular aid is to place a small memory, called a cache, right on the microprocessor itself. The cache holds those segments of a program that are most frequently used and thereby allows the processor to avoid calling on external memory chips much of the time. Some newer chips actually dedicate more transistors to the cache than they do to the processor itself. Future microprocessors will allot most resources to the cache to better bridge the speed gap.

The Holy Grail of computer design is an approach called parallel processing, which delivers the benefits of a single fast processor by engaging many inexpensive ones at the same time. In our analogy, we would go to a laundromat and use 20 washers and 20 dryers to do 20 loads simultaneously. Clearly, parallel processing is a costly solution for small workloads. And writing a program that can use 20 processors at once is much harder than distributing laundry to 20 washers. Indeed, the program must specify which instructions can be launched by which processor at what time.

Superscalar processing bears similarities to parallel processing, and it is more popular because the hardware automatically finds instructions that launch at the same time. But its potential processing power is not as large. If it were not so difficult to write the necessary programs, parallel processors could be made as powerful as one could afford. For the past 25 years, computer scientists have predicted that the programming problems will be overcome. In fact, parallel processing is practical for only a few classes of programs today.

In reviewing old articles, I have seen fantastic predictions of what computers would be like in 1997. Many stated that optics would replace electronics; computers would be built entirely from biological materials; the stored program concept would be discarded. These descriptions demonstrate that it is impossible to foresee what inventions will prove commercially viable and go on to revolutionize the computer industry. In my

CHARLES O'REAR



SILICON WAFERS today (*background*) are much larger but hold only about half as many individual chips as did those of the original microprocessor, the Intel 4004 (*foreground*). The dies can be bigger in part because the manufacturing process (*one stage shown in inset*) is cleaner.

Copyright 1997 Scientific American, Inc.

CHARLES O'REAR

CLEAN ROOMS, where wafers are made, are designed to keep human handling and airborne particles to a minimum. A single speck of dust can damage a tiny transistor.

career, only three new technologies have prevailed: microprocessors, random-access memory and optical fibers. And their impact has yet to wane, decades after their debut.

Surely one or two more inventions will revise computing in the next 25 years. My guess, though, is that the stored program concept is too elegant to be easily replaced. I believe future computers will be much like machines of the past, even if they are made of very different stuff. I do not think the microprocessor of 2020 will be startling to people from our time, although the fastest chips may be much larger than the very first wafer, and the cheapest chips may be much smaller than the original Intel 4004.

IRAMs and Picoprocessors

Pipelining, superscalar organization and caches will continue to play major roles in the advancement of microprocessor technology, and if hopes are realized, parallel processing will join them. What will be startling is that microprocessors will probably exist in everything from light switches to pieces of paper. And the range of applications these extraordinary devices will support, from voice recognition to virtual reality, will very likely be astounding.

Today microprocessors and memories are made on distinct manufacturing lines, but it need not be so. Perhaps in the near future, processors and memory will be merged onto a single chip, just as the microprocessor first merged the separate components of a processor onto a single chip. To narrow the processor-memory performance gap, to take advantage of parallel processing, to amortize the costs of the line and simply to make full use of the phenomenal number of transistors that can be placed on a single chip, I predict that the high-end microprocessor of 2020 will be an entire computer.

Let's call it an IRAM, standing for intelligent random-access memory, since most of the transistors on this merged chip will be devoted to memory. Whereas current microprocessors rely on hundreds of wires to connect to external memory chips, IRAMs will need no more than computer network connections and a power plug. All input-output devices will be linked to them via networks. If they need more memory, they will get more processing power and network connections as well, and vice versa—an arrangement that will keep the memory capacity and processor speed and network connectivity in balance. IRAMs are also the ideal building block for parallel processing. And because they would require so few external connections, these chips could be extraordinarily small. We may well see cheap "picoprocessors" that are smaller than the ancient Intel 4004. If parallel processing succeeds, this sea of transistors could also be used by multiple processors on a single chip, giving us a micromultiprocessor.

Today's microprocessors are more than 100,000 times faster than their 1950s ancestors, and when inflation is considered, they cost 1,000 times less. These extraordinary facts explain why computing plays such a large role in our world now. Looking ahead, microprocessor performance will easily keep doubling every 18 months through the turn of the century. After that, it is hard to bet against a curve that has outstripped all expectations. But it is plausible that we will see improvements in the next 25 years at least as large as those seen in the past 50. This estimate means that one desktop computer in 2020 will be as powerful as all the computers in Silicon Valley today. Polishing my crystal ball to look yet another 25 years ahead, I see another quantum jump in computing power.

The implications of such a breathtaking advance are limited only by our imaginations. Fortunately, the editors have asked others to ponder the possibilities, and I happily pass the baton on to them. SA

The Author

DAVID A. PATTERSON has taught since 1977 at the University of California, Berkeley, where he now holds the E. H. and M. E. Pardee Chair in Computer Science. He is a member of the National Academy of Engineering and is a fellow of both the Institute of Electrical and Electronic Engineers and the Association for Computing Machinery. He has won several teaching awards, co-authored five books and consulted for many companies, including Digital, Intel and Sun Microsystems. As a result of writing the original version of this article in 1995, he decided the following year to focus his research on intelligent random-access memory (IRAM). Some of the fruits of his work in this area are described on his World Wide Web site at http://iram.cs.berkeley.edu/

Further Reading

MICROPROCESSORS: FROM DESKTOPS TO SUPERCOMPUTERS. F. Baskett and J. L. Hennessy in *Science,* Vol. 261, pages 864–871; August 13, 1993.

COMPUTER ARCHITECTURE: A QUANTITATIVE APPROACH. Second edition. D. A. Patterson and J. L. Hennessy. Morgan Kaufmann Publishers, 1995.

COMPUTER ORGANIZATION AND DESIGN: THE HARDWARE/SOFTWARE INTERFACE. Second edition. D. A. Patterson and J. L. Hennessy. Morgan Kaufmann Publishers, 1997.

Follow the reference on the World Wide Web at http://cra.org:80/research.impact/ and look under "RISC" to learn more about the rapid rise in processor performance.

Copyright 1997 Scientific American, Inc.

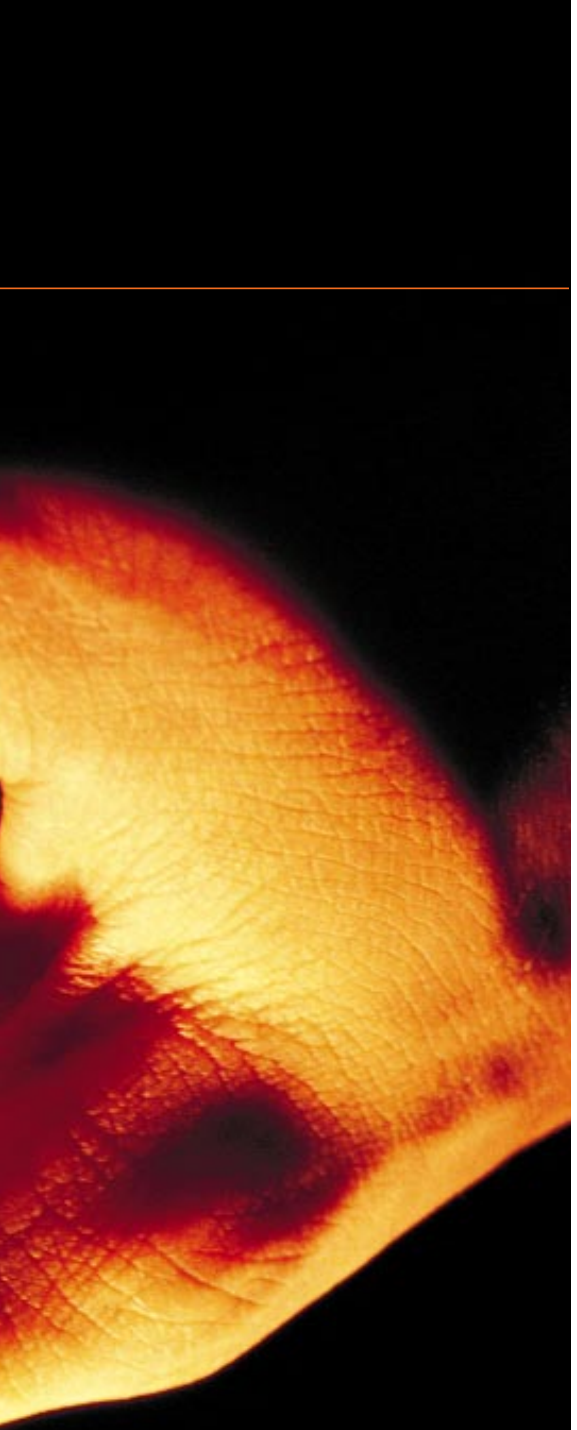
By tailoring the electrical properties of conducting polymers, researchers hope to render electronics a bit more organic

PLASTICS GET WIRED

by Philip Yam, *staff writer*

PLIABLE LIGHT shines from a polymer in this alphanumeric display made by UNIAX Corporation in Santa Barbara, Calif. Organic light-emitting diodes, or LEDs, should find applications soon and may form the basis of future lightweight screens.

Copyright 1997 Scientific American, Inc.



CHARLES O'REAR

Like many technological advances, the innovations in the field of conducting polymers began by accident. While attempting to make an organic polymer called polyacetylene in the early 1970s, Hideki Shirakawa of the Tokyo Institute of Technology mistakenly added 1,000 times more catalyst than the recipe called for. What he produced was a lustrous, silvery film that resembled aluminum foil but stretched like Saran Wrap—something that sounds more like a new and improved way to keep leftovers fresh than a potential breakthrough in materials science.

The substance appeared so unusual that when Alan G. MacDiarmid spied it, he wondered if it would be a candidate for his goal of making "synthetic metals"—nonmetallic substances that could transmit electricity. In 1977 Shirakawa joined MacDiarmid and Alan J. Heeger in their laboratory at the University of Pennsylvania to investigate this form of polyacetylene. After mixing in some iodine, the group found that the material's conductivity subsequently jumped by a factor of several million.

Durable, cheap, manufacturable and flexible, conducting polymers inspired visions of a future of transparent circuits, artificial muscle and electronic displays that conveniently roll up under the arm. Researchers have auditioned various demonstration devices, including components that could be useful for new displays, such as plastic transistors and light-emitting diodes (LEDs). Although such a future is about as dreamy as it gets, many investigators see broad marketing opportunities possible now—in antistatic coatings, electromagnetic shielding, lights for toys and microwave ovens, among others. Perhaps mundane, such applications are nonetheless promising enough that universities are collaborating with corporations, and scientists have initiated start-ups.

Although the pace of technological innovation has been impressively brisk, whether the materials will have an effect on commerce remains unclear. Firms are unlikely to invest in new equipment if the devices perform only marginally better than existing instruments. Polymer-based batteries, for instance, have a longer shelf life than do conventional ones, but they have penetrated the market in only a limited way. Flat-panel displays and LEDs made of organic substances face entrenched competition from existing inorganic liquid crystals and semiconductors.

Still, optimism pervades the field. Because plastic and electrical devices have become integral parts of the modern world, researchers are confident that at least some profitable uses will emerge. Conducting polymers constitute a radically novel market area, points out Ray H. Baughman of Allied-Signal in Morristown, N.J., who predicts confidently, "Fortunes are going to be made."

Polymers, the constituents of familiar plastic materials and synthetic fibers, are large organic molecules built out of smaller ones linked together in a long chain. Generally, they are insulators, because their molecules have no free electrons for carrying current. To make these substances conductive, workers exploit a technique familiar to the semiconducting industry: doping, or adding atoms with interesting electronic properties. The added atoms either give up some of their spare electrons to the polymer bonds or grab some electrons from the bonds (and thereby contribute positive charges called holes). In either case, the chain becomes electrically unstable. Applying a voltage can then send electrons scampering over the length of the polymer.

Since the Pennsylvania group's work, several kinds of polymers have been found to conduct electricity when doped. Besides polyacetylene, there are polypyrrole, polythiophene and polyaniline, to name just a few of the most commonly studied. Although scientists do not understand the precise physical mechanisms that enable polymers to conduct, the purity and particularly the arrangement of polymer chains seem to be crucial. By stretching polyacetylene, for instance, workers now routinely make the material conduct 50,000 amperes per volt per centimeter, up from 60 in the first reports. Some investigators have even managed to make polyacetylene con-

Copyright 1997 Scientific American, Inc.

duct about one quarter as well as copper.

Such developments are "extremely important for the whole conducting field," MacDiarmid says. "They exemplify how dedicated improvement in chemical and molecular structure can lead to enormous advances in the physical and electrical properties." Moreover, the conductivity is readily adjusted. "You can control the quality of the metallic state by controlling the structural order of the polymer," notes Arthur J. Epstein of Ohio State University.

Although other polymers are more conductive, polyaniline is emerging as the material of choice for many applications. As one of the oldest of synthetic organic polymers, its properties are well known. The substance—which resembles the plastic used in 35-millimeter photographic film—is easily made, it is stable in air and its electronic properties are readily customized. Most important, polyaniline is cheap—the most inexpensive conducting polymer around. In terms of geometry, it can also assume diverse incarnations, including thin films and patterned surfaces.

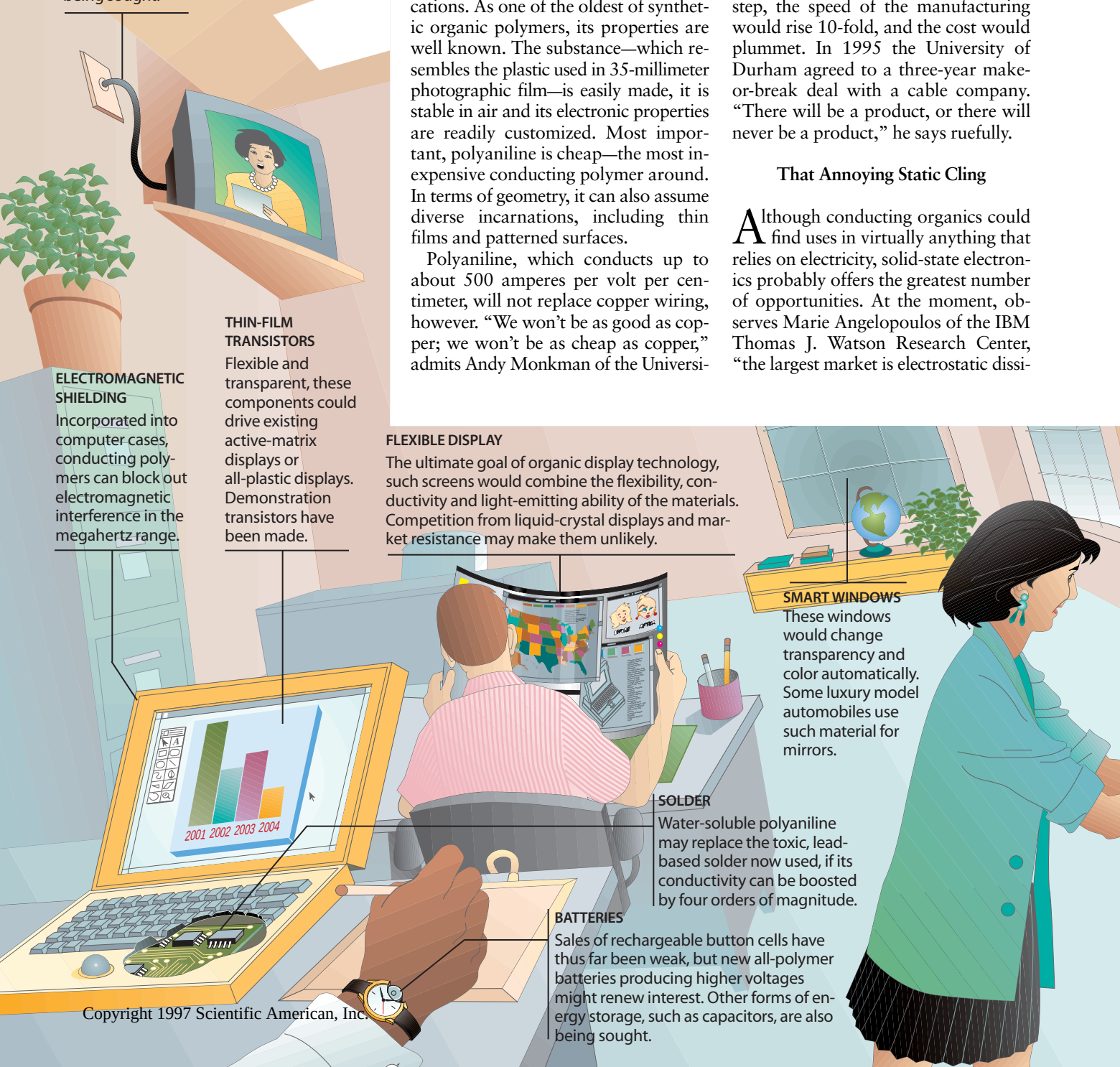
Polyaniline, which conducts up to about 500 amperes per volt per centimeter, will not replace copper wiring, however. "We won't be as good as copper; we won't be as cheap as copper," admits Andy Monkman of the University of Durham in England. Copper conducts 100,000 times as much current and costs half as much. Still, polyaniline's electrical performance is more than adequate for some applications, he insists: "The kinds of things we are going to replace are those that are complicated to manufacture, like braids on cable." Braids impart flexibility, permitting coaxial cable to wind around your living-room end table, for example, to reach the cable television box. But weaving copper wire into braids is a slow, laborious task, Monkman explains. If workers could extrude polymer braids and lay the insulation over the cable in a single step, the speed of the manufacturing would rise 10-fold, and the cost would plummet. In 1995 the University of Durham agreed to a three-year make-or-break deal with a cable company. "There will be a product, or there will never be a product," he says ruefully.

That Annoying Static Cling

Although conducting organics could find uses in virtually anything that relies on electricity, solid-state electronics probably offers the greatest number of opportunities. At the moment, observes Marie Angelopoulos of the IBM Thomas J. Watson Research Center, "the largest market is electrostatic dissi-

Conducting Plastics at Work

Displayed are some devices that might rely on electrically conducting organic materials in the near future.


COAXIAL CABLE
Polyaniline could replace copper in braided parts of the cable. Appropriate manufacturing techniques are now being sought.
THIN-FILM TRANSISTORS
Flexible and transparent, these components could drive existing active-matrix displays or all-plastic displays. Demonstration transistors have been made.
ELECTROMAGNETIC SHIELDING
Incorporated into computer cases, conducting polymers can block out electromagnetic interference in the megahertz range.
FLEXIBLE DISPLAY
The ultimate goal of organic display technology, such screens would combine the flexibility, conductivity and light-emitting ability of the materials. Competition from liquid-crystal displays and market resistance may make them unlikely.
SMART WINDOWS
These windows would change transparency and color automatically. Some luxury model automobiles use such material for mirrors.
SOLDER
Water-soluble polyaniline may replace the toxic, lead-based solder now used, if its conductivity can be boosted by four orders of magnitude.
BATTERIES
Sales of rechargeable button cells have thus far been weak, but new all-polymer batteries producing higher voltages might renew interest. Other forms of energy storage, such as capacitors, are also being sought.
2001 2002 2003 2004

Copyright 1997 Scientific American, Inc.

pation." Such charges are well known to wreak havoc on digital devices: estimates of electrostatic damage to electronic equipment alone exceed $15 billion yearly in the U.S., she notes.

Contemporary protective packaging, which relies on ionic salts or resins filled with metals or carbon, has some shortcomings. The conductivities of ionic materials tend to be low and unstable; metal is expensive and heavy; and carbon poses a contamination hazard because bits of it can slough off during shipment. Polymers should be easier to handle and able to dissipate electrostatic charges more efficiently. As a bonus, polyaniline coatings also happen to be highly transparent. In the summer of 1997 IBM began licensing the production of one such material, PanAquas.

The dissipative abilities of polymers also make them ideal for electromagnetic shielding. Such protection is necessary to keep electrical signals among components from overlapping—the reason airlines request that portable electronics be turned off during takeoff and landing. (The shielding would not benefit those concerned about the potential health effects of power lines, however, because the frequencies of the fields are much lower than these screens can block.) Incorporated into the plastic cases of electronic equipment, the polymers can guard against spurious signals, Epstein remarks. Conventional screening materials rely on impregnated bits of carbon or metal, which could harm the mechanical properties of the base material at any points that bend. Although proposals relying on polymers are still more costly than present solutions, conducting polymers could be adulterated with other substances, such as nylon, to reduce the expense.

Polymers could also be environmentally correct. IBM's PanAquas is soluble in water (ordinarily, the polymer must be processed with organic solvents). If Angelopoulos and her colleagues could increase the conductivity of the water-soluble polyaniline, the material could replace the lead-based solder used to connect electronics parts on a substrate. MacDiarmid explains that outdated equipment poses an environmental hazard and an economic nuisance: "In many parts of Europe the manufacturer must remove all lead-containing material from discarded printed circuit boards, which is one hell of a job."

The All-Plastic Transistor

The ultimate achievement in electronics application, however, would be a component fabricated out of polymers. Using ordinary circuit-printing

KARL GUDE
CAMOUFLAGE COATINGS
The U.S. military is considering coatings and fabrics that are blended with conducting polymers to spoof radar.
BIOLOGICAL SENSORS
Conductivity of polymer tags would change depending on exposure time above a threshold temperature and would be remotely read by a scanner. Sensors for aromas, enzymes and pesticides are now being used for quality control and safety analysis.
9:15
COOK 2
19:54
WASH
MEDICAL SUPPL
IL-12
ARTIFICIAL MUSCLE
Simple tweezers, made from strips of polymers with different conductivities, have been made; proposals call for muscle to consist of many strips working together.
LIGHT-EMITTING DIODES
Organic LEDs can glow in all colors, even white. Such lights will be usable for appliance panels, watches and toys.
ANTISTATIC MATERIALS
Polymers that dissipate electrostatic charge may be used in flooring, mouse pads and computer-chip packaging.

Copyright 1997 Scientific American, Inc.

techniques, Francis Garnier of the CNRS Molecular Materials Laboratory in Thiais, France, did just that, creating the first all-polymer circuit element: a transistor. Constructed around a short-chain molecule called sexithiophene, the thin-film field-effect transistor was fully flexible. Twisting, rolling and bending (even at right angles) had no effect on the electrical characteristics of the device.

Although widely regarded as an impressive bit of engineering, Garnier's organic transistor would not stand a chance against silicon. Computers made from the plastic material would operate at less than one thousandth the speed of existing ones crafted of crystalline silicon, which permits electrons to move faster.

But there is an application that does not need fast electronics: video displays. Currently amorphous silicon (that is, silicon in its noncrystalline form) is used in such circuitry because it is much less expensive to process than crystals are and can be laid on different substrates, such as glass. Garnier's transistor runs at just about the speed of circuits made from amorphous silicon, and he feels the requisite video-rate performance is easily within reach.

An organic semiconducting transistor would be a boon to manufacturers of liquid-crystal displays (LCDs), the approach that dominates research into flat-panel technology. Existing screens seal liquid crystals, made from various kinds of organic substances, between two glass plates; a fluorescent tube illuminates the crystals from behind. In so-called passive displays, the pixels (cells containing the liquid crystals) are controlled by voltages applied along all the rows and columns. In active-matrix displays, which offer greater contrast and resolution, each pixel is individually controlled by a thin-film transistor.

Therein lies the cost. A 20-inch, full-color active-matrix display contains more than two million pixels. Unfortunately, a few malfunctioning ones are sufficiently distracting to the sensitive human eye to ruin the image. "The percentage of flat panels rejected is very high," Garnier states. That failure rate drives up the price of the displays that make it to market.

Organic circuits might ease the strain on corporate wallets because they should be easier to make, especially in large sizes. The circuitry can be fabricated at lower temperatures and is less sensitive to the presence of impurities during processing, which should lower production costs. Moreover, organics could make it possible to fabricate entirely new types of displays. Manufacturers should be able to tune the properties of the polymers, controlling their flexibility and even their transparency. See-through electronics would permit a direct-view, heads-up display on windshields and helmets, obviating the need to reflect images onto a viewing glass, as is now done.

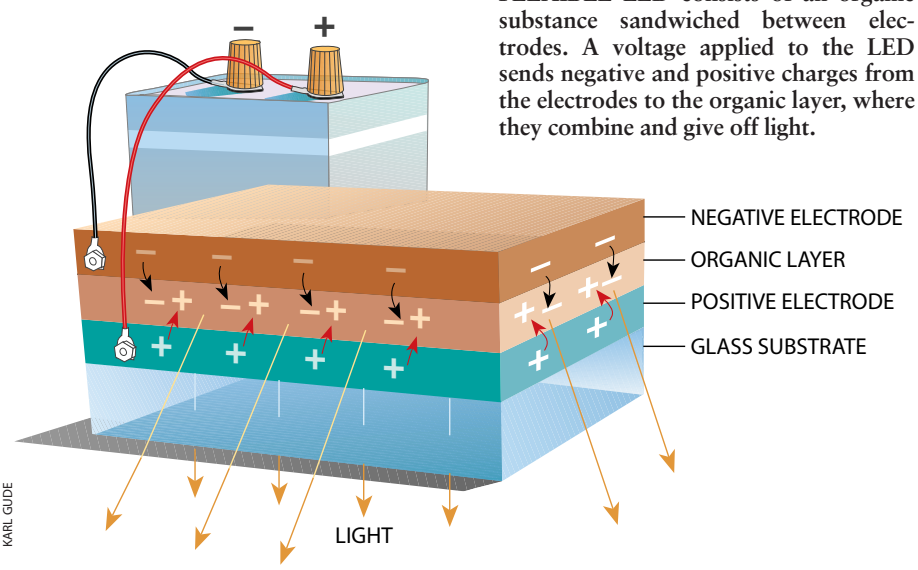
Shines in the Dark

Conducting organics could also be used as the light sources in displays, not just in the controlling circuitry. Indeed, lightweight, robust displays have been one of the most widely publicized, pie-in-the-sky applications. But as a first step researchers are aiming for a more modest, albeit lucrative, use—light-emitting diodes. These little glowing indicators decorate innumerable electronic gizmos, and their sales add up to hundreds of millions of dollars a year in the U.S. alone, according to the Semiconductor Industries Association in Santa Clara, Calif.

At present, LEDs are constructed from an inorganic semiconducting material, often gallium arsenide. Two layers, each doped to have different electrical characteristics, are interconnected and act as positive and negative electrodes. When electricity passes through the materials, one electrode gives off electrons, the other, positively charged holes (spaces that electrons would normally occupy). The negative and positive charges meet at the junction of the substances, where they combine and give off light. The color of the light depends on the properties of the semiconductor and dopant; those producing red and green light are the easiest to make.

Organic LEDs promise to make the manufacture of these lights much cheaper, mostly by reducing the number of contacts and interconnections. Conventional LEDs must be spliced together to be used in displays on such devices as microwave ovens, alarm clocks and videocassette recorders. Each LED cannot be crafted larger than the gallium arsenide crystal wafers can be grown, and modern technology limits the size to no more than about six inches, measured diagonally. To make a large display, then, LEDs must be individually mounted and wired—a difficult task considering that one reasonably sized letter in a typical display takes 35 LEDs. In contrast, organic films can be laid over practically unlimited extents. In addition, the starting materials for organics are more economical than those for conventional LEDs.

Ching W. Tang and his colleagues at Eastman Kodak are by far leading the way in bringing organic-based LEDs to market. (The rather undescriptive term for the approach they have adopted—"small molecule"—distinguishes it from work using polymers, which are much longer.) In 1987 Tang reported that a small crystalline organic molecule of 8-hydroxyquinoline aluminum (Alq) would give off green and orange light. Since then, workers found they could elicit all colors of the spectrum by varying the thin-film organic layer. More-



KARL GUDE

FLEXIBLE LED consists of an organic substance sandwiched between electrodes. A voltage applied to the LED sends negative and positive charges from the electrodes to the organic layer, where they combine and give off light.

Copyright 1997 Scientific American, Inc.

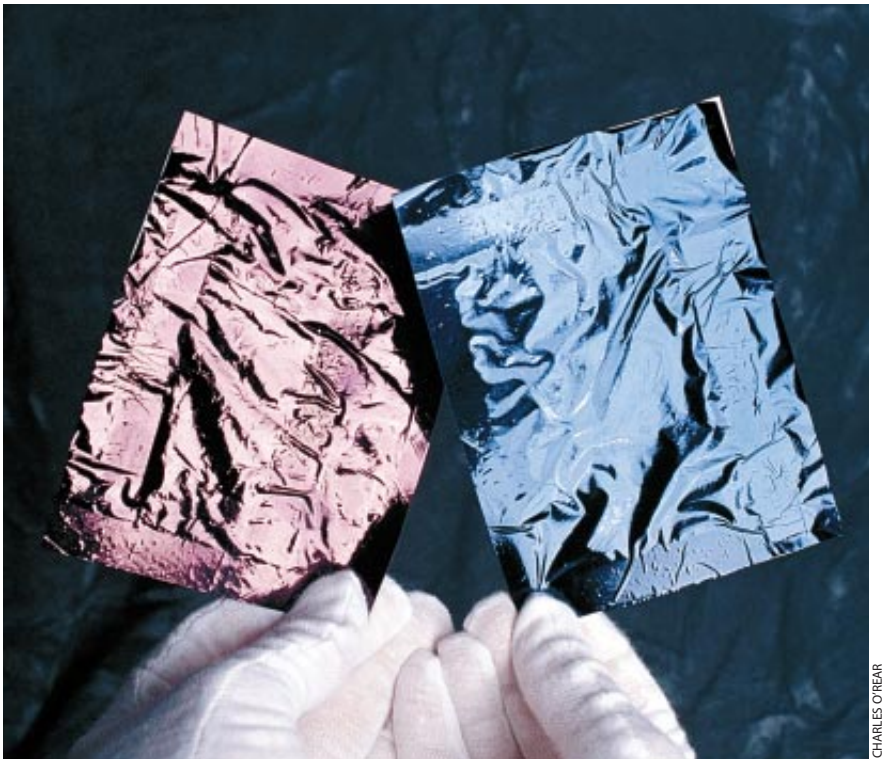
over, the organic LEDs can, in lumens per watt, burn as efficiently as a household lightbulb and can last 10 times longer—in other words, more than 10,000 hours.

"The efficiency is extremely attractive. With [components that have] a 10,000-hour lifetime," Tang says, "you can seriously consider display applications, particularly in portable devices." At the moment, small-molecule LEDs are not ready to replace liquid-crystal displays in flat screens—their performance is still too poor. Yet it is adequate for dot-matrix displays in electronic organizers and microwave oven panels, for instance, and that will do for now.

High-end displays are not completely out of reach. In the autumn of 1996 Samson A. Jenekhe and his colleagues at the University of Rochester built a polymer diode consisting of two organic plastic layers. By varying the voltage applied across the two layers, the researchers were able to vary the color of the emitted light from red through yellow to green.

In March 1997 Jenekhe and his colleagues were finally able to produce an organic plastic material that emitted blue light, and in the summer of 1997 researchers at Princeton University and the University of Southern California created a bright, red-green-blue organic LED whose color can be independently controlled. Before these accomplishments can be developed into a product, however, several significant issues must be resolved, including the stability of these organic materials over long periods. In the meantime, researchers are investigating other potential uses for the plastic light emitters, including backlights for liquid-crystal displays.

All the devices built so far, though, have been too dim and inefficient. One solution for increasing the brightness and efficiency may be an alternative architecture. An approach that has shown some promise was unveiled recently at Bell Laboratories (now part of Lucent Technologies), where Ananth Dodabalapur and his colleagues constructed electroluminescent devices by sandwiching layers of Alq and inert material between two reflecting surfaces. Structured this way, the layers conform to the physics of a Fabry-Perot cavity—the basic structure of most lasers. The emissive Alq sends out light that bounces back and forth, amplifying until it leaks out one end. (This type of microcavity yielded true lasing in 1996.) Because the light emerges from only one end, more of it reaches the viewer, unlike the light from conventional diode structures, which leaks wastefully in all directions.



CHARLES O'REAR

POLYMER SHEETS made of polyaniline appear as a lustrous pink (*left*) until doped with iodine, which makes the substance conduct and colors it blue (*right*). Weigong Zheng of the University of Pennsylvania prepared the material.

The potentially higher efficiency may also boost the longevity. Current that is not transformed into light becomes waste heat, which hastens a diode's demise. Because a microcavity LED would require less current for the same amount of light, it should in principle last longer.

Polymer Lights

Other investigators are trying to develop LEDs made from polymers instead of small organic molecules. The most widely used polymers are poly-p-phenylenevinylene, or PPV for short, and its derivatives. Richard H. Friend of the Cavendish Laboratory at the University of Cambridge and his associates discovered the green-yellow glow of PPV in 1990. By combining that material with electrodes made from other polymers or from flexible metal backings (like the foil that keeps supermarket nachos fresh), researchers have produced flexible LEDs that give off 2.5 lumens per watt. Driven at 10 volts, the light is about as bright as the fluorescent lamp in a liquid-crystal display. By varying the chemical makeup of PPV, they have also teased the full range of colors out of the devices. In 1996 several researches showed that PPV can even lase.

So far, however, polymer LEDs have plenty of drawbacks. "Lifetime issues are clearly key to making this curiosity into a business," remarks Heeger, now at the University of California at Santa Barbara. Most polymer LEDs burn for only a few hundred hours, but 2,000 to 10,000 hours is desirable. The main cause is inefficiency. The polymer LEDs convert no more than 4 percent of the current sent through them into light; the rest is transformed into waste heat. Hence, the diode can shine quite brightly, but the high voltage necessary to achieve that intensity comes at the price of faster breakdown.

Improved processing might extend PPV's life; during manufacturing, unwanted reactions can create defects on the polymer chain, which interfere with PPV's ability to glow. Shelf life is also a drawback; at the moment, PPV diodes last only several months in storage because they are unstable in air, reacting with oxygen and water vapor. Better packaging might help.

Still, polymer LEDs are close to being sufficiently bright and efficient for

Copyright 1997 Scientific American, Inc.

some limited applications. Cambridge Display Technology, which Friend helped to found, has licensed its technology to Heeger's company, UNIAX Corporation in Santa Barbara. In conjunction with the Dutch giant Philips Electronics and the chemical maker Hoechst AG, by the end of 1997 UNIAX plans to make small displays, about 2.5 by five centimeters in size and containing between 2,000 and 3,000 pixels. The polymers might also find use as lights in toys, watches and promotional novelties.

Even if lifetime issues are resolved, polymer LEDs may never really see the light of day, not so long as the small-molecule, Alq-based LEDs surpass them in performance. Japan has focused virtually all its attention on the small-molecule lights. Pioneer Electronics, for instance, used Kodak's Alq technology to demonstrate the LEDs in alphanumeric displays containing up to some 16,000 pixels that can burn for 5,000 hours. What keeps hope alive in the polymer crowd is the potential for cheaper manufacturing. Polymer LEDs extracted from solutions of chemicals may be easier to make than small-molecule LEDs, which are deposited in a high vacuum onto substrates.

Who Wants Wallpaper That Glows?

Whether any new kind of LED—small-molecule or polymer—emerges on a large scale depends on manufacturability. "Almost certainly at a cost, anything can be done," Friend notes. "The question is whether these things are going to be cheap." More to the point, existing technology is quite adequate. As indicator lights, conventional LEDs cost only pennies. As backlights, standard fluorescent lights are excellent sources, comments Lewis J. Rothberg, formerly at Bell Labs, now at the University of Rochester. For polymer products, he says, "the competition is going to be harsh."

The color capability of organics could also be irrelevant. Why would you need a rainbow of hues if you just want to know if your amplifier is on? More broadly, does a market for a large, roll-up display truly exist? That question still has no clear answer. "People have a vision of carrying around a view graph," Rothberg says. "I don't know if the public is going to want that."

There is some justification for skepticism. The first commercial products incorporating conducting polymers were actually made a few years ago. In the late 1980s the Japanese companies Bridgestone and Seiko commercialized a rechargeable button-cell battery that used polyaniline for one electrode and lithium for the other. Milliken and Company, a textile manufacturer based in South Carolina, developed Contex, a fabric that consists of common synthetics interwoven with the conducting polymer polypyrrole. It just so happened that the conductivity of the resulting fabric was perfect for "spoofing" radar—that is, interfering with detection by making it appear that the signals were going right through empty space. It has an advantage over the military's existing radar camouflage nets, which rely on incorporated carbon fibers, in that it has no gaps in its signal absorption.

"The question is whether these things are going to be cheap," says Richard H. Friend of the University of Cambridge.

Yet sales of these early products proved disappointing. Although the polymer-based battery had a longer shelf life than did lead-acid or nickel-cadmium cells, the technology never took off. Heeger explains that the advantage, though real, was not substantial enough to convince investors to set up completely new manufacturing plants. (There might be room for specialized applications, though. For instance, workers at Johns Hopkins University made an all-plastic, rechargeable battery in early 1997. Flexible and light, it can produce up to three volts—sufficient for some satellite and battlefield equipment, for which weight is a factor.)

Commercialization of Contex was perhaps even more discouraging. "We were approved as a vendor for the A-12 bomber," remarks Hans H. Kuhn of Milliken, "but the bomber was never built." Although sobered, Kuhn is hoping that the army's interest in camouflage nets could revive appeal for the material.

Another product that has proved disappointing is the electronic nose, which works because odor molecules can alter the resistance of conducting polymers. The U.K.-based firm Aromascan, the first to commercialize electronic noses in 1994, has posted mounting losses; in 1996 it reached $1.8 million as commercial interest in the noses—for quality control and scent analysis, among other uses—has slipped since the introduction of the devices.

Even conducting polymers that have loyal customers may not be financially worthwhile for a big corporation. Before IBM's PanAquas antistatic spray coating, Allied-Signal offered an analogous product named Versacon—the main difference being that Versacon was a dispersible powder rather than a solution and therefore may not have been as effective or as transparent. At the time, several companies considered Versacon advantageous and incorporated it into such products as paints and coatings. Yet Allied has abandoned production; the volume of sales was simply too low. "The major problems for wide applications remain cost and reliability," says Epstein of Ohio State.

That does not faze the pioneers of conducting polymers, especially because possibilities beyond electronics are conceivable. Epstein has a patent on a technique that uses the polymers to form "hidden joints." Polyaniline in powder form can be sprinkled on two pieces of plastic that need to be joined. The conducting powder can absorb the energy from ordinary microwave ovens to heat and fuse the pieces, making them as strong as one.

Baughman and MacDiarmid have made plastic electromechanical mechanisms. Two polymers with different conductivities would change their linear dimensions when current flows through them, much as the metallic strips in thermostats do under varying temperatures. The polymers would undergo more dramatic changes in size using much less electricity than conventional piezoelectric or electrostatic actuators, Baughman says. More than just high-tech tweezers, several microactuators coupled together could function as artificial muscle.

Certainly there is no shortage of imagination, and such immediate uses as the dissipation of static charge and the shielding of electromagnetic fields are clearly viable. But stiff competition from present-day devices and marketing considerations may jeopardize hopes of having a portable roll-up display to take on the commute to work. The newspaper may have to do for a while. SA

Copyright 1997 Scientific American, Inc.

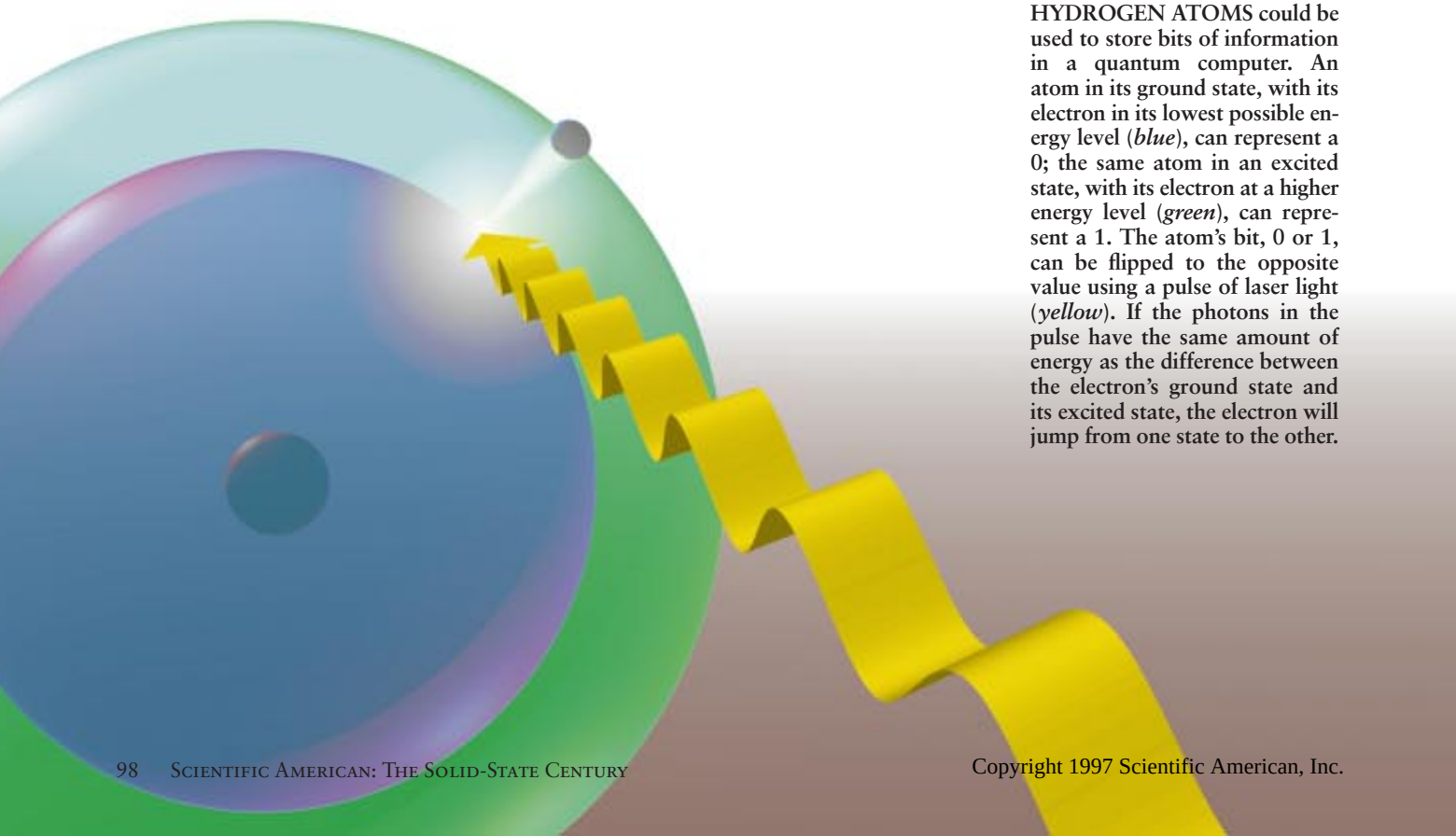
Quantum-mechanical computers, if they can be constructed, will do things no ordinary computer can

Quantum-Mechanical Computers

by Seth Lloyd

Every two years for the past 50, computers have become twice as fast while their components have become half as big. Circuits now contain wires and transistors that measure only one hundredth of a human hair in width. Because of this explosive progress, today's machines are millions of times more powerful than their crude ancestors. But explosions do eventually dissipate, and integrated-circuit technology is running up against its limits.

Advanced lithographic techniques can yield parts $^1/_{100}$ the size of what is currently available. But at this scale—where bulk matter reveals itself as a crowd of individual atoms—integrated circuits barely function. A tenth the size again, the individuals assert their identity, and a single defect can wreak havoc. So if computers are to become much smaller in the future, new technology must replace or supplement what we now have.



HYDROGEN ATOMS could be used to store bits of information in a quantum computer. An atom in its ground state, with its electron in its lowest possible energy level (*blue*), can represent a 0; the same atom in an excited state, with its electron at a higher energy level (*green*), can represent a 1. The atom's bit, 0 or 1, can be flipped to the opposite value using a pulse of laser light (*yellow*). If the photons in the pulse have the same amount of energy as the difference between the electron's ground state and its excited state, the electron will jump from one state to the other.

Copyright 1997 Scientific American, Inc.

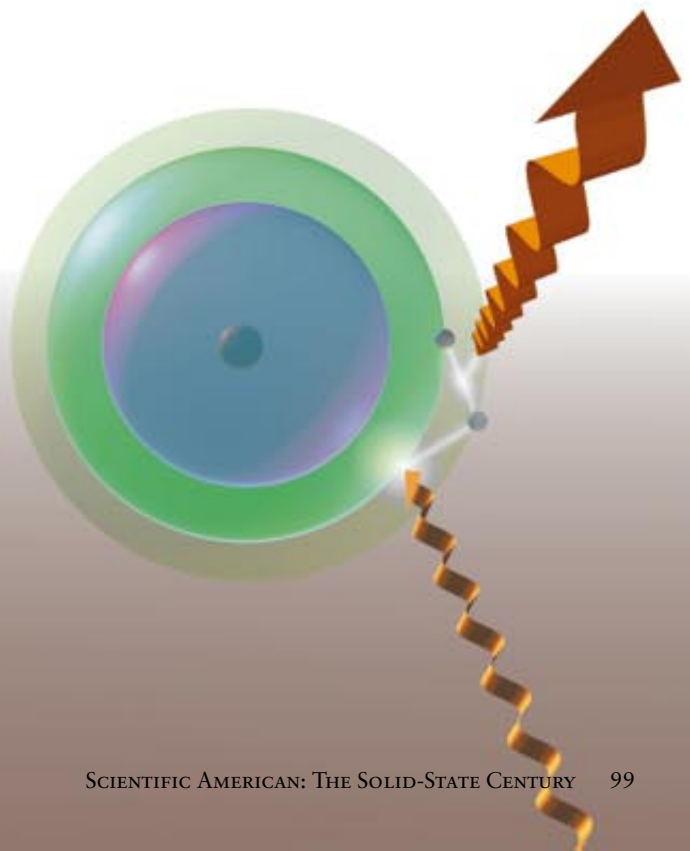
Several decades ago pioneers such as Rolf Landauer and Charles H. Bennett, both at the IBM Thomas J. Watson Research Center, began investigating the physics of information-processing circuits, asking questions about where miniaturization might lead: How small can the components of circuits be made? How much energy must be used up in the course of computation? Because computers are physical devices, their basic operation is described by physics. One physical fact of life is that as the components of computer circuits become very small, their description must be given by quantum mechanics.

In the early 1980s Paul Benioff of Argonne National Laboratory built on Landauer and Bennett's earlier results to show that a computer could in principle function in a purely quantum-mechanical fashion. Soon after, David Deutsch of the Mathematical Institute at the University of Oxford and other scientists in the U.S. and Israel began to model quantum-mechanical computers to find out how they might differ from classical ones. In particular, they wondered whether quantum-mechanical effects might be exploited to speed computations or to perform calculations in novel ways.

By the middle of the decade, the field languished for several reasons. First, all these researchers had considered quantum computers in the abstract instead of studying actual physical systems—an approach that Landauer faulted on many counts. It also became evident that a quantum-mechanical computer might be prone to errors and have trouble correcting them. And apart from one suggestion, made by Richard Feynman of the California Institute of Technology, that quantum computers might be useful for simulating other quantum systems (such as new or unobserved forms of matter), it was unclear that they could solve mathematical problems any faster than their classical cousins.

In the past few years, the picture has changed. In 1993 I described a large class of familiar physical systems that might act as quantum computers in ways that avoid some of Landauer's objections. Peter W. Shor of AT&T Bell Laboratories has demonstrated that a quantum computer could be used to factor large numbers—a task that can foil the most powerful of conventional machines. And in 1995, workshops at the Institute for Scientific Interchange in Turin, Italy, spawned many designs for constructing quantum circuitry. More recently, H. Jeff Kimble's group at Caltech and David J. Wineland's team at the National Institute of Standards and Technology have built some of these prototype parts, whereas David Cory of the Massachusetts Institute of Technology and Isaac Chuang of Los Alamos National Laboratory have demonstrated simple versions of my 1993 design. This article explains how quantum computers might be assembled and describes some of the astounding things they could do that digital computers cannot.

BORIS STAROSTA

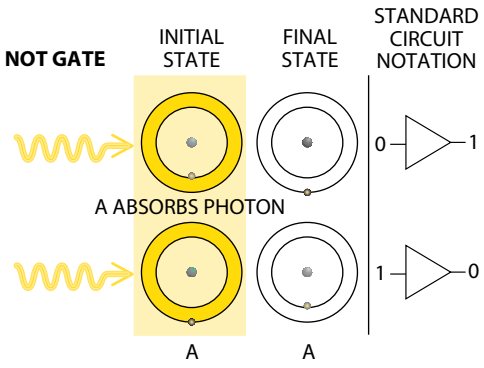


READING the bit an atom stores is done using a laser pulse having the same amount of energy as the difference between the atom's excited state, call it E_1, and an even higher, less stable state, E_2. If the atom is in its ground state, representing a 0, this pulse has no effect. But if it is in E_1, representing a 1, the pulse pushes it to E_2. The atom will then return to E_1, emitting a telltale photon.

Copyright 1997 Scientific American, Inc.

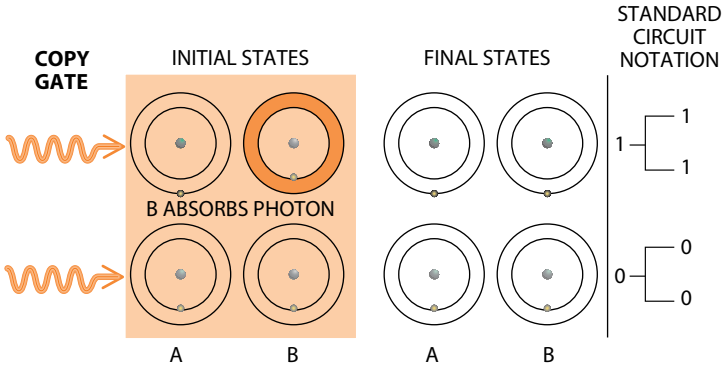
Quantum Logic Gates

Logic gates are devices that perform elementary operations on bits of information. The Irish logician George Boole showed in the 19th century that any complex logical or arithmetic task could be accomplished using combinations of three simple operations: NOT, COPY and AND. In fact, atoms, or any other quantum system, can perform these operations. —*S.L.*



NOT involves nothing more than bit flipping, as the notation above shows: if *A* is 0, make it a 1, and vice versa. With atoms, this can be done by applying a pulse whose energy equals the difference between *A*'s ground state (its electron is in its lowest energy level, shown as the inner ring) and its excited state (shown as the outer ring). Unlike conventional NOT gates, quantum ones can also flip bits only halfway.

COPY, in the quantum world, relies on the interaction between two different atoms. Imagine one atom, *A,* storing either a 0 or 1, sitting next to another atom, *B,* in its ground state. The difference in energy between the states of *B* will be a certain value if *A* is 0, and another value if *A* is 1. Now apply a pulse of light whose photons have an energy equal to the latter amount. If the pulse is of the right intensity and duration and if *A* is 1, *B* will absorb a photon and flip (*top row*); if *A* is 0, *B* cannot absorb a photon from the pulse and stays unchanged (*bottom row*). So, as in the diagram below, if *A* is 1, *B* becomes 1; if *A* is 0, *B* remains 0.



MICHAEL GOODMAN

Let's face it, quantum mechanics is weird. Niels Bohr, the Danish physicist who helped to invent the field, said, "Anyone who can contemplate quantum mechanics without getting dizzy hasn't properly understood it." For better or worse, quantum mechanics predicts a number of counterintuitive effects that have been verified experimentally again and again. To appreciate the weirdness of which quantum computers are capable, we need accept only a single strange fact called wave-particle duality.

Wave-particle duality means that things we think of as solid particles, such as basketballs and atoms, behave under some circumstances like waves and that things we normally describe as waves, such as sound and light, occasionally behave like particles. In essence, quantum-mechanical theory sets forth what kind of waves are associated with what kind of particles, and vice versa.

The first strange implication of wave-particle duality is that small systems such as atoms can exist only in discrete energy states. So when an atom moves from one energy state to another, it absorbs and emits energy in exact amounts, or "chunks," called photons, which might be considered the particles that make up light waves.

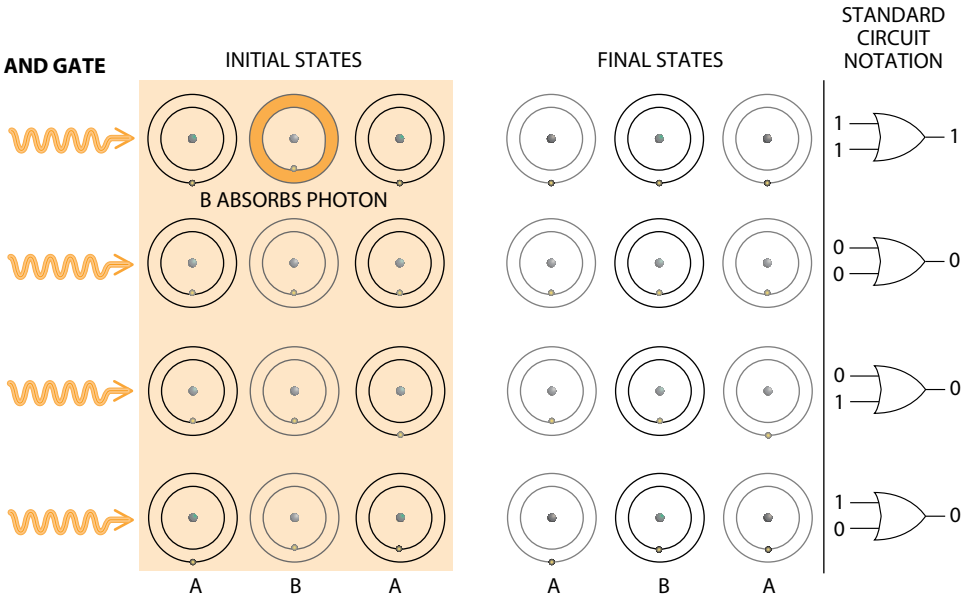
A second consequence is that quantum-mechanical waves, like water waves, can be superposed, or added together. Taken individually, these waves offer a rough description of a given particle's position. When two or more such waves are combined, though, the particle's position becomes unclear. In some weird quantum sense, then, an electron can sometimes be both here and there at the same time. Such an electron's location will remain unknown until some interaction (such as a photon bouncing off the electron) reveals it to be either here or there but not both.

When two superposed quantum waves behave like one wave, they are said to be coherent; the process by which two coherent waves regain their individual identities is called decoherence. For an electron in a superposition of two different energy states (or, roughly, two different positions within an atom), decoherence can take a long time. Days can pass before a photon, say, will collide with an object as small as an electron, exposing its true position. In principle, basketballs could be both here and there at once as well (even in the absence of Michael Jordan). In practice, however, the time it takes for a photon to bounce off a ball is too brief for the eye or any instrument to detect. The ball is simply too big for its exact location to go undetected for any perceivable amount of time. Consequently, as a rule only small, subtle things exhibit quantum weirdness.

Quantum Information

Information comes in discrete chunks, as do atomic energy levels in quantum mechanics. The quantum of information is the bit. A bit of information is a simple distinction between two alternatives—no or yes, 0 or 1, false or true. In digital computers, the voltage between the plates in a capacitor represents a bit of information: a charged capacitor registers a 1 and an uncharged capacitor, a 0. A quantum computer functions by matching the familiar discrete character of digital information processing to the strange discrete character of quantum mechanics.

Copyright 1997 Scientific American, Inc.



AND also depends on atomic interactions. Imagine three atoms, *A*, *B* and *A*, sitting next to one another. The difference in energy between the ground and excited states of *B* is a function of the states of the two *A*'s. Suppose *B* is in its ground state. Now apply a pulse whose energy equals the difference between the two states of *B* only when the atom's neighboring *A*'s are both 1. If, in fact, both *A*'s are 1, this pulse will flip *B* (*top row*); otherwise it will leave *B* unchanged (*all other rows*).

Indeed, a string of hydrogen atoms can hold bits as well as a string of capacitors. An atom in its electronic ground state could encode a 0 and in an excited state, a 1. For any such quantum system to work as a computer, though, it must be capable of more than storing bits. An operator must be able to load information onto the system, to process that information by way of simple logical manipulations and to unload it. That is, quantum systems must be capable of reading, writing and arithmetic.

Isidor Isaac Rabi, who was awarded the Nobel Prize for Physics in 1944, first showed how to write information on a quantum system. Applied to hydrogen atoms, his method works as follows. Imagine a hydrogen atom in its ground state, having an amount of energy equal to E_0. To write a 0 bit on this atom, do nothing. To write a 1, excite the atom to a higher energy level, E_1. We can do so by bathing it in laser light made up of photons having an amount of energy equal to the difference between E_1 and E_0. If the laser beam has the proper intensity and is applied for the right length of time, the atom will gradually move from the ground state to the excited state, as its electron absorbs a photon. If the atom is already in the excited state, the same pulse will cause it to emit a photon and go to the ground state. In terms of information storage, the pulse tells the atom to flip its bit.

What is meant here by gradually? An oscillating electrical field such as laser light drives an electron in an atom from a lower energy state to a higher one in the same way that an adult pushes a child on a swing higher and higher. Each time the oscillating wave comes around, it gives the electron a little push. When the photons in the field have the same energy as the difference between E_0 and E_1, these pushes coincide with the electron's "swinging" motion and gradually convert the wave corresponding to the electron into a superposition of waves having different energies.

The amplitude of the wave associated with the electron's ground state will continuously diminish as the amplitude of the wave associated with the excited state builds. In the process, the bit registered by the atom "flips" from the ground state to the excited state. When the photons have the wrong frequency, their pushes are out of sync with the electron, and nothing happens.

If the right light is applied for half the time it takes to flip the atom from 0 to 1, the atom is in a state equal to a superposition of the wave corresponding to 0 and the wave corresponding to 1, each having the same amplitudes. Such a quantum bit, or qubit, is then flipped only halfway. In contrast, a classical bit will always read either 0 or 1. A half-charged capacitor in a conventional computer causes errors, but a half-flipped qubit opens the way to new kinds of computation.

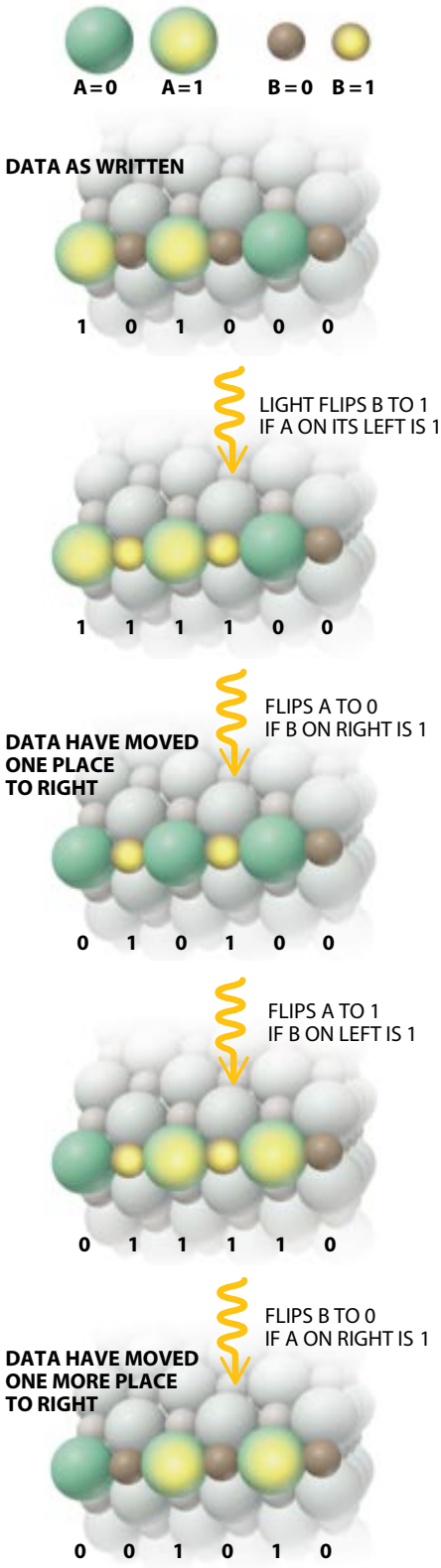
Reading bits from a quantum system is similar to flipping them. Push the atom to an even higher, less stable energy state, call it E_2. Do so by subjecting the atom to light having an energy equal to the difference between E_1 and E_2: if the atom is in E_1, it will be excited to E_2 but decay rapidly back to E_1, emitting a photon. If the atom is already in the ground state, nothing happens. If it is in the "half-flipped" state, it has an equal chance of emitting a photon and revealing itself to be a 1 or of not emitting a photon, indicating that it is a 0. From writing and reading information in a quantum system, it is only a short step to computing.

Quantum Computation

Electronic circuits are made from linear elements (such as wires, resistors and capacitors) and nonlinear elements (such as diodes and transistors) that manipulate bits in different ways. Linear devices alter input signals individually. Nonlinear devices, on the other hand, make the input signals passing through them interact. If your stereo did not contain nonlinear transistors, for example, you could not change the bass in the music it plays. To do so requires some coordination of the information coming from your compact disc and the information coming from the adjustment knob on the stereo.

Circuits perform computations by way of repeating a few simple linear and nonlinear tasks over and over at great speed. One such task is flipping a bit, which is equivalent to the logical operation called NOT: true becomes false, and false be-

Copyright 1997 Scientific American, Inc.



comes true. Another is COPY, which makes the value of a second bit the same as the first. Both those operations are linear, because in both the output reflects the value of a single input. Taking the AND of two bits—another useful task—is a nonlinear operation: if two input bits are both 1, make a third bit equal to 1 as well; otherwise make the third bit a 0. Here the output depends on some interaction between the inputs.

The devices that execute these operations are called logic gates. If a digital computer has linear logic gates, such as NOT and COPY gates, and nonlinear ones as well, such as AND gates, it can complete any logical or arithmetic task. The same requirements hold for quantum computers. Artur Ekert, working with Deutsch and Adriano Barenco at Oxford, and I have shown independently that almost any nonlinear interaction between quantum bits will do. Indeed, provided a quantum computer can flip bits, any nonlinear quantum interaction enables it to perform any computation. Hence, a variety of physical phenomena might be exploited to construct a quantum computer.

In fact, all-purpose quantum logic gates have been around almost as long as the transistor! In the late 1950s, researchers managed to perform simple two-bit quantum logic operations using particle spins. These spins—which are simply the orientation of a particle's rotation with respect to some magnetic field—are, like energy levels, quantized. So a spin in one direction can represent a 1 and in the other, a 0. The researchers took advantage of the interaction between the spin of the electron and the spin of the proton in a hydrogen atom; they set up a system in which they flipped the proton's spin only if the electron's spin represented a 1. Because these workers were not thinking about quantum logic, they called the effect double resonance. And yet they used double resonance to carry out linear NOT and COPY operations.

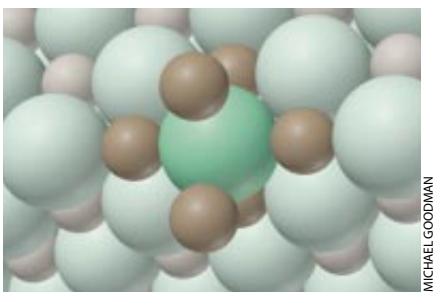
Since then, Barenco, David DiVincenzo of IBM, Tycho Sleator of New York University and Harald Weinfurter of the University of Innsbruck have demonstrated how, by flipping proton and electron spins only partway, double resonance can be used to create an AND gate as well. Such quantum logic gates, wired together, could make a quantum computer.

A number of groups have recently constructed quantum logic gates and proposed schemes for wiring them together. A particularly promising development has come from Caltech: by concentrating photons together with a single atom in a minute volume, Kimble's group has enhanced the usually tiny nonlinear interaction between photons. The result is a quantum logic gate: one photon bit can be flipped partway when another photon is in a state signifying 1. Quantum "wires" can be constructed by having single photons pass through optical fibers or through the air, in order to ferry bits of information from one gate to another.

An alternative design for a quantum logic circuit has been proposed by J. Ignacio Cirac of the University of Castilla-La Mancha in Spain and Peter Zoller of the University of Innsbruck. Their scheme isolates qubits in an ion trap, effectively insulating them from any unwanted external influences. Before a bit were processed, it would be transferred to a common register, or "bus." Specifically, the information it contained would be represented by a rocking motion involving all the ions in the trap. Wineland's group at NIST has taken the first step in realizing such a quantum computer, performing both linear and nonlinear operations on bits encoded by ions and by the rocking motion.

In an exciting theoretical development under experimental investigation at Caltech, Cirac, Zoller, Kimble and Hideo Mabuchi have shown how the photon and ion-trap schemes for quantum computing might be combined to create a "quantum Internet" in which photons are used to shuttle qubits coherently back and forth between distant ion traps.

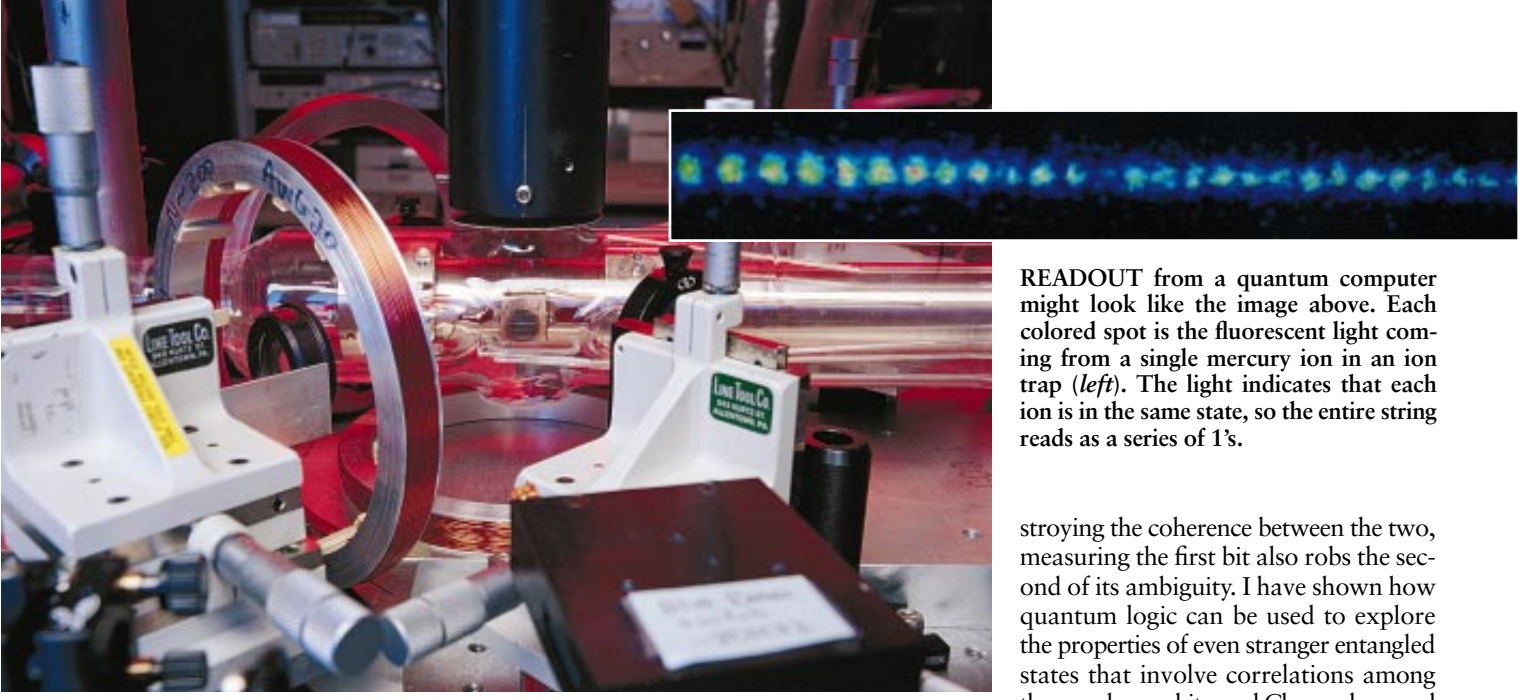
Although their methods can in princi-



MICHAEL GOODMAN

SALT CRYSTAL could be made to compute by acting on pairs of neighboring ions. Flip the bit held by each *B* if the *A* on its left stores a 1; then flip each *A* if the *B* on its right is 1. This moves the information from each *A* to the *B* on its right. Now, using the same tactics, move the information from each *B* to the *A* on its right. The process allows a line of atoms to act as a quantum "wire." Because a crystal can carry out these "double resonance" operations simultaneously in all directions with every neighboring ion (*bottom, right*), the crystal can mimic the dynamics of any system and so serves as a general-purpose quantum analog computer.

Copyright 1997 Scientific American, Inc.



READOUT from a quantum computer might look like the image above. Each colored spot is the fluorescent light coming from a single mercury ion in an ion trap (*left*). The light indicates that each ion is in the same state, so the entire string reads as a series of 1's.

GEOFFREY WHEELER (*left*); NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (*right*)

ple be scaled up to tens or hundreds of quantum bits, the Caltech and NIST groups have performed quantum logic operations on just two bits (leading some wags to comment that a two-bit microprocessor is just a two-bit microprocessor). In 1997, however, Neil A. Gershenfeld of M.I.T., together with Chuang of Los Alamos, showed that my 1993 method for performing quantum computing using the double resonance methods described above could be realized using nuclear spins at room temperature. The same result was obtained independently by M.I.T.'s Cory, working with Amr Fahmy and Timothy F. Havel of Harvard Medical School. With conventional magnets of the kind used to perform magnetic resonance imaging, Chuang and Cory both succeeded in performing quantum logic operations on three bits, with the prospect of constructing 10-bit quantum microprocessors in the near future.

Thus, as it stands, scientists can control quantum logic operations on a few bits, and in the near future, they might well do quantum computations using a few tens or hundreds of bits. How can this possibly represent an improvement over classical computers that routinely handle billions of bits? In fact, even with one bit, a quantum computer can do things no classical computer can. Consider the following. Take an atom in a superposition of 0 and 1. Now find out whether the bit is a 1 or a 0 by making it fluoresce. Half of the time, the atom emits a photon, and the bit is a 1. The other half of the time, no photon is emitted, and the bit is a 0. That is, the bit is a random bit—something a classical computer cannot create. The random-number programs in digital computers actually generate pseudorandom numbers, using a function whose output is so irregular that it seems to produce bits by chance.

Multiparticle Quantum States

Imagine what a quantum computer can do with two bits. Copying works by putting together two bits, one with a value to be copied and one with an original value of 0; an applied pulse flips the second bit to 1 only if the first bit is also 1. But if the value of the first bit is a superposition of 0 and 1, then the applied pulse creates a superposition involving both bits, such that both are 1 or both are 0. Notice that the final value of the first bit is no longer the same as it was originally—the superposition has changed.

In each component of this superposition, the second bit is the same as the first, but neither is the same as the original bit. Copying a superposition state results in a so-called entangled state, in which the original information no longer resides in a single quantum bit but is stored instead in the correlations between qubits. Albert Einstein noted that such states would violate all classical intuition about causality. In such a superposition, neither bit is in a definite state, yet if you measure one bit, thereby putting it in a definite state, the other bit also enters into a definite state. The change in the first bit does not *cause* the change in the second. But by virtue of destroying the coherence between the two, measuring the first bit also robs the second of its ambiguity. I have shown how quantum logic can be used to explore the properties of even stranger entangled states that involve correlations among three and more bits, and Chuang has used magnetic resonance to investigate such states experimentally.

Our intuition for quantum mechanics is spoiled early on in life. A one-year-old playing peekaboo knows that a face is there even when she cannot see it. Intuition is built up by manipulating objects over and over again; quantum mechanics seems counterintuitive because we grow up playing with classical toys. One of the best uses of quantum logic is to expand our intuition by allowing us to manipulate quantum objects and play with quantum toys such as photons and electrons.

The more bits one can manipulate, the more fascinating the phenomena one can create. I have shown that with more bits, a quantum computer could be used to simulate the behavior of any quantum system. When properly programmed, the computer's dynamics would become exactly the same as the dynamics of some postulated system, including that system's interaction with its environment. And the number of steps the computer would need to chart the evolution of this system over time would be directly proportional to the size of the system.

Even more remarkable, if a quantum computer had a parallel architecture, which could be realized through the exploitation of the double resonance between neighboring pairs of spins in the atoms of a crystal, it could mimic any quantum system in real time, regardless of its size. This kind of parallel quantum computation, if possible, would give a huge speedup over conventional methods. As Feynman noted, to simulate a quantum system on a classical computer generally requires a number of steps

Copyright 1997 Scientific American, Inc.

that rises exponentially both with the size of the system and with the amount of time over which the system's behavior is tracked. In fact, a 40-bit quantum computer could re-create in little more than, say, 100 steps, a quantum system that would take a classical computer, having a trillion bits, years to simulate.

What can a quantum computer do with many logical operations on many qubits? Start by putting all the input bits in an equal superposition of 0 and 1, each having the same magnitude. The computer then is in an equal superposition of of all possible inputs. Run this input through a logic circuit that carries out a particular computation. The result is a superposition of all the possible outputs of that computation. In some weird quantum sense, the computer performs all possible computations at once. Deutsch has called this effect "quantum parallelism."

Quantum parallelism may seem odd, but consider how waves work in general. If quantum-mechanical waves were sound waves, those corresponding to 0 and 1—each oscillating at a single frequency—would be pure tones. A wave corresponding to a superposition of 0 and 1 would then be a chord. Just as a musical chord sounds qualitatively different from the individual tones it includes, a superposition of 0 and 1 differs from 0 and 1 taken alone: in both cases, the combined waves interfere with each other.

A quantum computer carrying out an ordinary computation, in which no bits are superposed, generates a sequence of waves analogous to the sound of "change ringing" from an English church tower, in which the bells are never struck simultaneously and the sequence of sounds follows mathematical rules. A computation in quantum-parallel mode is like a symphony: its "sound" is that of many waves interfering with one another.

Factoring could be an easy task for a quantum computer.

Shor of Bell Labs has shown that the symphonic effect of quantum parallelism might be used to factor large numbers very quickly—something classical computers and even supercomputers cannot always accomplish. Shor demonstrated that a quantum-parallel computation can be orchestrated so that potential factors will stand out in the superposition the same way that a melody played on violas, cellos and violins an octave apart will stand out over the sound of the surrounding instruments in a symphony. Indeed, his algorithm would make factoring an easy task for a quantum computer, if one could be built. Because most public-key encryption systems—such as those protecting electronic bank accounts—rely on the fact that classical computers cannot find factors having more than, say, 100 digits, quantum-computer hackers would give many people reason to worry.

Whether or not quantum computers (and quantum hackers) will come about is a hotly debated question. Recall that the quantum nature of a superposition prevails only so long as the environment refrains from somehow revealing the state of the system. Because quantum computers might still consist of thousands or millions of atoms, only one of which need be disturbed to damage quantum coherence, it is not clear how long interacting quantum systems can last in a true quantum superposition. In addition, the various quantum systems that might be used to register and process information are susceptible to noise, which can flip bits at random.

Shor and Andrew Steane of Oxford have shown that quantum logic operations can be used to construct error-correcting routines that protect the quantum computation against decoherence and errors. Further analyses by Wojciech Zurek's group at Los Alamos and by John Preskill's group at Caltech have shown that quantum computers can perform arbitrarily complex computations as long as only one bit in 100,000 is decohered or flipped at each time step.

It remains to be seen whether the experimental precision required to perform arbitrarily long quantum computations can be attained. To surpass the factoring ability of current supercomputers, quantum computers using Shor's algorithm might need to follow thousands of bits over billions of steps. Even with the error correction, because of the technical problems described by Landauer, it will most likely prove rather difficult to build a computer to perform such a computation. To surpass classical simulations of quantum systems, however, would require only tens of bits followed for tens of steps, a more attainable goal. And to use quantum logic to create strange, multiparticle quantum states and to explore their properties is a goal that lies in our current grasp. SA

The Author

SETH LLOYD is the Finmeccanica Career Development Professor in the mechanical engineering department at the Massachusetts Institute of Technology. He received his first graduate degree in philosophy from the University of Cambridge in 1984 and his Ph.D. in physics from the Rockefeller University in 1988. He has held postdoctoral positions at the California Institute of Technology and at Los Alamos National Laboratory, and since 1989 he has been an adjunct assistant professor at the Santa Fe Institute in New Mexico.

Further Reading

Quantum-Mechanical Models of Turing Machines That Dissipate No Energy. Paul Benioff in *Physical Review Letters,* Vol. 48, No. 23, pages 1581–1585; June 7, 1982.

Quantum Theory: The Church-Turing Principle and the Universal Quantum Computer. David Deutsch in *Proceedings of the Royal Society of London,* Series A, Vol. 400, No. 1818, pages 97–117; 1985.

A Potentially Realizable Quantum Computer. Seth Lloyd in *Science,* Vol. 261, pages 1569–1571; September 17, 1993.

Algorithms for Quantum Computation: Discrete Logarithms and Factoring. Peter W. Shor in *35th Annual Symposium on Foundations of Computer Science: Proceedings*. Edited by Shafi Goldwasser. IEEE Computer Society Press, 1994.

Quantum Computations with Cold Trapped Ions. J. I. Cirac and P. Zoller in *Physical Review Letters,* Vol. 74, No. 20, pages 4091–4094; May 15, 1995.

Copyright 1997 Scientific American, Inc.

An item that looks like one of today's laptops will be a portal into a personal network with phenomenal organizational skills

THE FUTURE OF THE PC

by Brad Friedlander and Martyn Roetter

"I think there is a world market for maybe five computers."
—*Thomas J. Watson, IBM chairman, 1943*

Prediction is fraught with peril. The sudden arrival in our culture of tens of millions of personal computers, however, necessarily makes one ponder what the near future will bring. Although the personal computer lets users perform many different useful or entertaining tasks, such as writing letters, playing games, sending e-mail or surfing the World Wide Web, it may already be—in its current incarnation—in its last days.

Rather than interact solely with a personal computer, users will engage a "personal network," which is already evolving from today's personal computer and its related devices. The current PC's capabilities barely scratch the surface of what the future may have in store. The changeover could be well along within a decade.

The typical personal network (PN) will consist of one or more computing devices in communication with one another, via a mix of permanent and transient communications links. At least one of the devices in a PN will be recognizable as a descendant of today's ubiquitous PC. Resembling current high-end laptops, this device will be the primary portal by which the user will access the network. (The portal will also function as a conventional PC when it is disconnected from the rest of the PN.)

In fact, rudimentary PNs already exist. Recently one of us (Friedlander) was working at home and had two laptop computers and a desktop computer linked together in a tiny local-area network (LAN). I was using two of the machines while moving information to them from the third. My wife happened by during this episode. She watched for a moment and then asked if I knew how silly the whole scene looked. I had to admit that I did. Yet that little contretemps made me realize that the PN already exists, albeit flimsily.

All the connections of my little network were out in the open, but tomorrow's fully realized PN will be transparent, as invisible to the user as the networks that provide electricity to every outlet in the home. And, for the most part, the network will require hardly any intervention to operate.

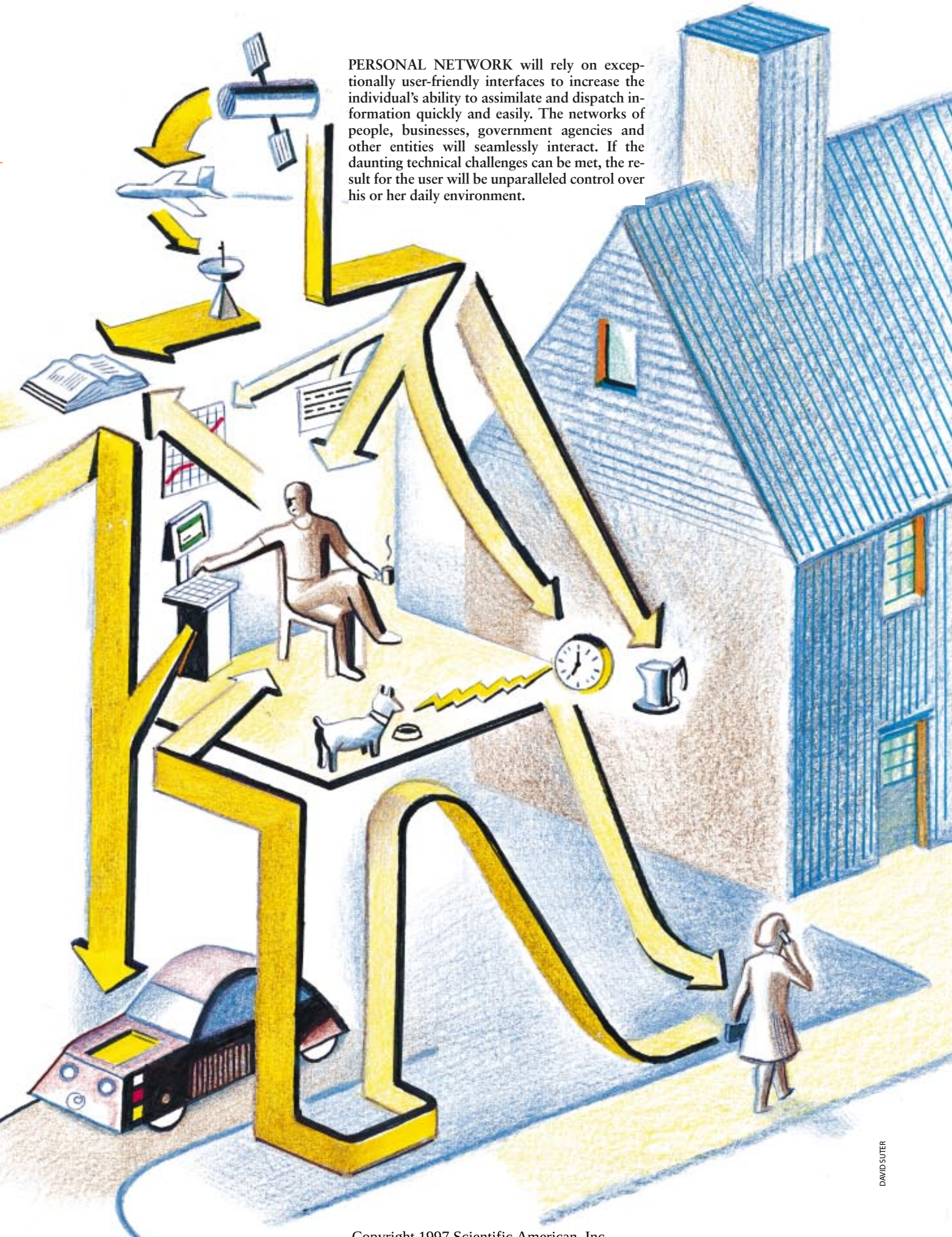
Whereas the advent and success of the PC can be legitimately classified as a revolution, the PN will be the product of evolution. PNs will use the same basic architecture that computers have employed since the days of Thomas J. Watson's slightly flawed prediction. That architecture, like the architecture of PCs now, will consist of a central processing unit (CPU), memory and input/output (I/O) devices. The CPU (for example, a Motorola PowerPC chip or an Intel Pentium II) does the brainwork, manipulating data. Memory, which includes random-access memory, or RAM, stores the information, such as the data that the CPU is currently using. Meanwhile the I/O devices literally do just that, providing the means for information to flow into and out of the computer. Typical I/O devices include those designed for interaction, such as the keyboard, screen and mouse; those involved with storage—for example, disks and CD-ROMs (compact disc, read-only memory); and those that serve as communications devices (modems and Ethernet LAN cards).

In the conventional PC, the CPU, memory and I/O devices are connected via one or more buses, circuits that provide the communications link making it possible for the various components of the PC to share data.

In the PN, the network itself is, in effect, the bus: the network architecture implies data flow in any and all directions, with the individual computers fully aware of the network and its constituents. Like the members of a baseball team interacting to make a play, member devices in the network will often work together. Just as players come to bat individually, however, each network member will operate some of the time without linking to other devices in the network.

Some of these individual devices within the PN will be dedicated to you, its owner, whereas others will be shared with other users, via their PNs. Again, the primary means for you to access the PN will be a portable computer resembling a high-end laptop. This unit will be dedicated to

Copyright 1997 Scientific American, Inc.

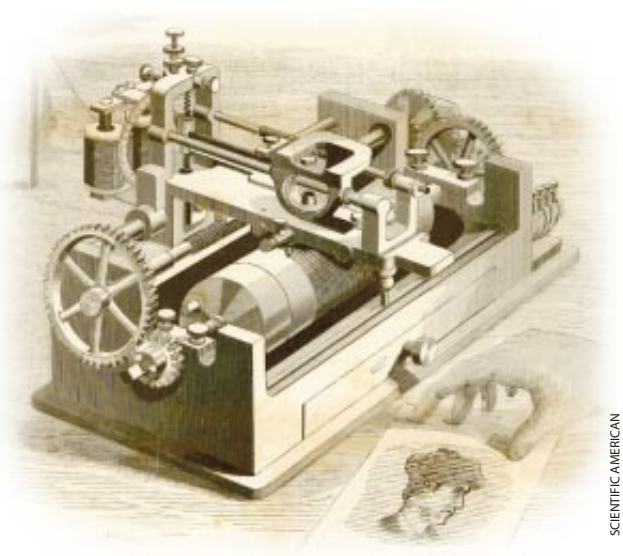


PERSONAL NETWORK will rely on exceptionally user-friendly interfaces to increase the individual's ability to assimilate and dispatch information quickly and easily. The networks of people, businesses, government agencies and other entities will seamlessly interact. If the daunting technical challenges can be met, the result for the user will be unparalleled control over his or her daily environment.

DAVID SUTER

Copyright 1997 Scientific American, Inc.

IBM



SCIENTIFIC AMERICAN

EARLY INFORMATION PROCESSORS were large and often difficult to use. IBM's seminal 701 computer took up a whole room (*left*); the "electro-artograph," an early facsimile machine, could work only with a photographic negative, which was engraved at the receiving end. Tomorrow's personal computers will also span enormous spaces, but in virtual fashion. High-speed processors, advanced software and wireless communications will come together to make connections transparent.

you. Other parts of the network could include computing devices embedded in your home, in your appliances or in your car. Because other family members will no doubt have their own PNs, some of these individual devices could be part of multiple PNs. Various servers, which are relatively powerful computers that anchor the network by performing computationally intensive services for less powerful computers on the network, will manage the PNs and their data stores. These servers will be shared among a number of PNs, but the individual PNs will remain logically separated.

One of the qualities of the PN that will make it so versatile will be its dynamic and seamless linking to the PNs of family members, friends or groups with a common interest, as well as to shared environments, such as the home or office. A combination of wire-based and wireless communications will make it possible to establish links to other PNs anywhere in the world, whenever you need to.

A Day in the Life

So much, for now, for the skeletal details of the personal network. A scenario describing a network in action illustrates its power and attraction. Imagine a Wednesday morning in the autumn of 2007. You awake to the sounds of your favorite music, playing at 7:00 A.M., a full half-hour later than usual. Your personal network has let you sleep late, based on your appointment calendar for the day. During your last moments of sleep, the PN collected information from various news outlets, assigning them priorities based on personal-interest profiles you had created. It turned up the temperature in part of your home (waiting that extra half an hour today) and instructed your coffeemaker to brew up a fresh pot.

When you sat down to drink your first cup of coffee, a thin but meaty morning newspaper was hot off your printer, specifically tailored to appeal to your concerns. Via voice communication and a large, flat-screen display, your PN notifies you that one of its computing devices has failed. Not to worry, however: the network has already submitted a repair order, and, by the way, no data were lost, thanks to automatic backups in other network devices. You could read the paper on the display, but you still prefer to actually hold it in your hands. As you check last night's box scores for the two teams you follow, the PN reminds you that your wife is away on business. "Yes," you respond as you continue to read and sip, "I noticed as soon as I woke up." The PN ignores your flippant tone to further remind you that she will be flying home tonight.

Shortly before it is time to leave for work, your PN starts to replicate information onto your laptop. This includes items stored on a server within your PN, such as the newspaper, your calendar and the documents that you revised since arriving home last night. In contrast to today's laptop, which houses all your data, the PN laptop carries only copies of data from servers on your PN and on your office and other networks. Thus, the data are decentralized and can be updated whenever necessary. This structure is part of what provides the ability to recover information fully and easily when a single computing device fails.

As you check your tie in the mirror, the PN notifies the navigation computer in your car of your departure time and destination this morning. The network switches the laptop over to sleep mode to conserve battery life and to allow you to disconnect it from its link. As you leave the house, the car engine roars to life. Your vocal command unlocks the door.

DAVID SUTER

IT KNOWS WHEN YOU'RE AWAKE and many other aspects about your private life. Moreover, it will convey that information only to those who need it and who have permission to have it. Encryption technology will keep personal information hidden from those outside the individual's sphere of sanctioned interactions.

Copyright 1997 Scientific American, Inc.

As you back out of the driveway, the navigation computer gets wind of a traffic snarl, informs you and offers an alternate route, estimates your new arrival time, notifies your office staff of your situation and reschedules your early meetings.

When you arrive at the office, your PN, via the laptop, synchronizes with the office network, bringing relevant items, such as those documents you revised last night, up to date. As you settle in at your desk, your PN links with your wife's PN to get her arrival schedule and to make a dinner reservation at her favorite restaurant. (You'll still get points for being so thoughtful.)

The day passes, highlighted by a video teleconference and a few old-fashioned telephone calls and meetings. As you prepare to leave work, your PN replicates the information you need to take home, synchronizes calendars and learns that your wife's flight is one hour late. It reschedules the dinner reservation and notifies your car, and your evening begins.

Getting There from Here

Although many activities in this imaginary day are achievable now, the protagonist would need to be a highly skilled computer user and would have to interact frequently, if not constantly, with each computer in the PN to make this scenario work. Furthermore, the equipment would cost a small fortune. All these factors should rapidly change.

Other examples of nascent PNs already in existence go a bit further than the three computers in the living room illustration. Many people have palmtop devices, such as the U.S. Robotics Pilot, and synchronize data between these and their laptop or desktop. Mobile users frequently connect into a LAN at work to exchange information with other people and to copy any changed files from their laptop to the office desktop. And some homes already have LANs that include both personal and shared resources—separate computers for each parent and child with printers and scanners that can be used by anyone on the network.

For the type of sophisticated network described in our workday setting to come into being, however, various technological hurdles still need to be cleared. Certainly, processing power must continue to improve dramatically and the cost of that power must continue to fall. But these improvements are among the lowest of the hurdles.

One of the most critical missing pieces of technology is the user interface. The PN will have to interact with people in ways that we humans find natural, so as to be attractive to those who may have far fewer computer skills than contemporary users do. The voice identification in our scenario is a prime example of the coming changes, as are speech recognition and speech synthesis by the computer. Users should ultimately be able to interact with their PNs simply by talking and listening to them. (Remember how the crew of the *Enterprise* spoke with the shipboard computer on *Star Trek*?)

Enabling machines to understand us enough of the time to be useful still represents a formidable challenge. Captain Kirk sounds quite different from Scotty. Speech patterns may even vary for a single person, depending on his or her mood or physical health. Nevertheless, systems already exist that allow users some vocal control of PCs, including entry of data, text and commands. The active vocabulary can reach tens of thousands of words. Depending on aural aptitude, the prices for these systems range from about $100 to as high as $1,500.

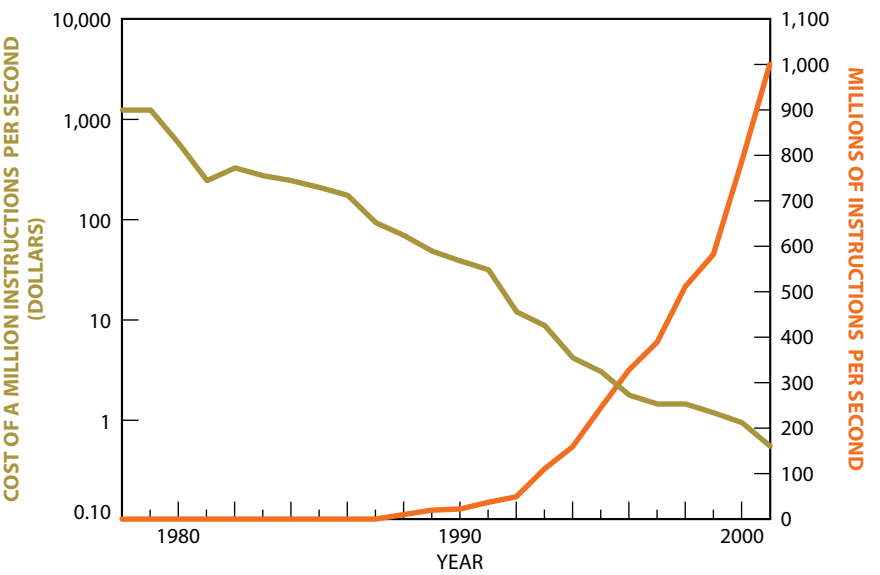
Speech recognition is based on identification of phonemes, the smallest acoustical components of language. (Spoken English includes roughly 80 phonemes.) Human speech is sampled by the computer, with the digitized audio signals compared with those on file. Some leeway must be available to accommodate variations in individuals' pitches and word duration. In addition, the probability of a particular word following another will clue in the computer to the likelihood of a match.

Machine-to-Machine Bonding

In addition to communication between human and machine, the machines must be able to talk to one another reliably, especially if the PN is to have the capability to interact easily with other people's PNs. Means of communication include physical links between machines, such as in home LANs; infrared links among collocated devices in very close proximity; short-range radio links, as in cordless phones; long-range radio links, as in cellular phones; and cable or high-speed connections to the rest of the world. Both wired and wireless communications require improvement for the PN to become a reality.

Wireless communications techniques are clearly central to PN mobile computing. Users must be able to transmit and access information regardless of their or its location and without a physical connection. For in-room communication, infrared systems can already provide links of up to four megabits per second (Mbps) over distances of about one meter. Wide-area usage requires radio links.

The challenges of wireless communications are far greater than those posed by wired links, whether copper or fiber-optic cables. Basic communication pa-



SOURCES: Intel; Semico Research Corporation

INCREASES IN PROCESSING SPEED and decreases in costs for those ever higher speeds will be major factors that allow the development of personal networks.

JARED SCHNEIDMAN DESIGN

Copyright 1997 Scientific American, Inc.

rameters, such as transmission speed, error rate, reliability and delay, can change substantially and suddenly during the course of a single wireless communication. Such instability is rooted in the variability of radio-frequency noise and in the signal attenuation that can arise from natural sources, such as storms, or interference from various electronic devices. Mobility throws yet more unknowns into the equation. For example, radio-wave propagation can be suddenly interrupted when a building, vehicle or some other large object comes between the transmitting and receiving devices. Even the presence or absence of foliage may affect transmissions, causing seasonal variations in performance.

For these and other reasons, the transmission speeds available in wireless communications tend to be much slower than those over wired links. The implications of these limitations and characteristics of wireless communications for software are considerable. While reading, you may be momentarily distracted by the sound of a car alarm, but you have no trouble returning to your place and the flow of information. Applications that are expected to function effectively in an environment involving wireless channels must likewise be able to handle variable communications performance and recover from unexpected but inevitable interruptions, keeping track of partial or incomplete results.

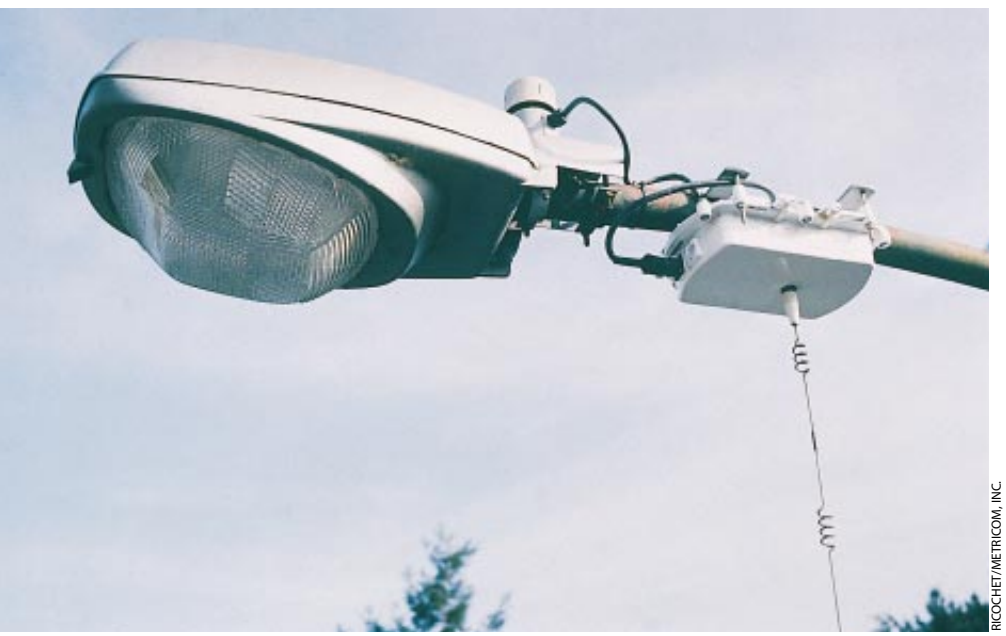
Despite these disadvantages, the great utility of wireless data communications has kept interest high. Among the options now being considered or developed are the use of existing cellular telephone channels for data, as well as dedicated wireless data systems. Effective wireless data-transmission speeds (up to 10,000 bits per second) are significantly slower than modems over dial-up wired telephone lines (up to 33,600 bits per second). Data speeds over modern digital cellular systems, however, such as the European GSM (Global System for Mobile Communications), are beginning to reach these higher levels.

Wireless data speeds currently lag well behind those of various modern data services that users can obtain from phone companies, such as so-called frame relay lines or the Integrated Services Digital Network (ISDN). Both can offer rates in the megabits-per-second range. But the next generation of mobile wireless communications systems, known as IMT 2000 (International Mobile Telecommunications), is aiming at speeds of up to two million bits per second. As the name implies, IMT 2000 is being developed by the international telecommunications community for introduction at the turn of the century.

Longer-term research and development efforts are being directed toward what is known as wireless asynchronous transfer mode (ATM) to reach rates of 155 million bits per second or even higher. This system has the further advantage of compatibility with the anticipated spread of ATM-based services in wired telecommunications. And although today's mobile wireless services transmit at frequencies of 800 to 900 megahertz and two gigahertz, the Federal Communications Commission has designated frequencies in the tens of gigahertz range for future use as the lower frequencies become crowded.

In anticipation of this bandwidth availability, high-speed satellite services are already being planned for shortly after 2000. One key project is the Teledesic multisatellite array, using several hundred satellites in low-earth orbits. Boeing, Microsoft founder Bill Gates and cellular-telephone business tycoon Craig McCaw are all investing in this technology. The Teledesic array will downlink at 18 gigahertz and uplink at 28 gigahertz. It therefore should be able to provide coverage at up to gigabit-per-second speeds even in the most remote parts of the globe. Stanley would be able to call Dr. Livingstone, who could then send a digital map image directly to Stanley's laptop. ("High data-rate satellite link, I presume?")

Having all these data flying around necessarily raises privacy issues. One wants to ensure that the PN follows orders solely from its owner and that transmissions reach only those people or devices one wants to reach (as Newt Gingrich, Prince Charles and other notables recently discovered, when their supposedly private cellular phone conversations became newspaper headlines). In other words, the personal network must be secure.



RICOCHET/METRICOM, INC.

RADIO TRANSMITTERS inconspicuously placed on light poles a quarter mile apart have already established a wireless communications network over sections of Seattle, San Francisco and Washington, D.C. A wireless modem in communication with these devices, which in turn interface with wired access points, enables users of the network (built by Metricom) to access information far from telephone lines.

Open, Says Me

Access to PN devices will be governed by biometrics, which is the use of physiological features, such as the pattern of a voice, fingerprint or iris, to permit activation (for example, voice recognition allows you and only you to enter your car). Classic controls, such as passwords, will also supplement biometrics.

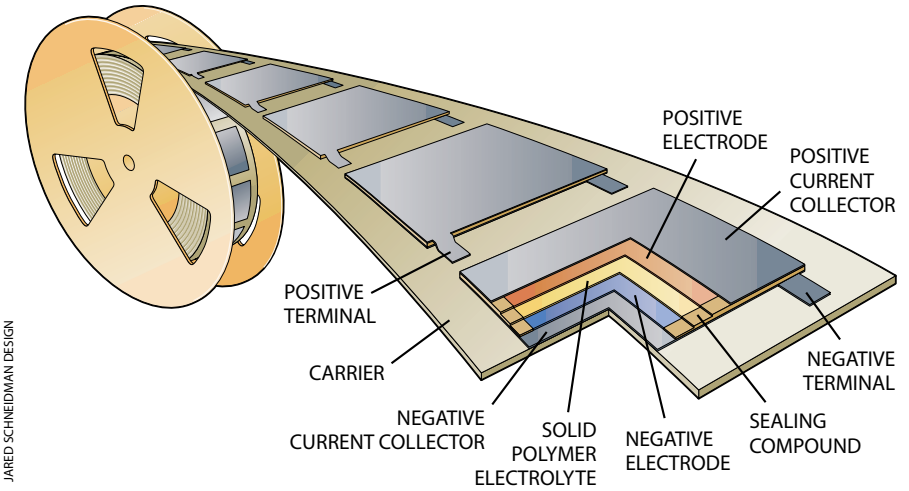
The core of security, however, will be public-key cryptography, a mathematically based system that will be used to secure owner control over the PN and allow for privacy both within the network and when linking with other net-

Copyright 1997 Scientific American, Inc.

works. (The encryption systems known as RSA and PGP are examples of public-key cryptography.)

Public-key cryptography works by assigning each person two different keys, which are actually two related numbers. Both keys are many digits in length and would be impossible for most people to memorize. They are typically stored on your computer's hard drive or on a disk. The public key is available to anyone who wants to send you a secure message. The private key, however, is known only to you. As the name implies, public-key cryptography can ensure encryption and therefore privacy. A sender encrypts a message to you—say, a credit-card number—using your public key. The message can be deciphered, however, only with both the public and the private key—and only you have the latter.

Another critical technology is software. Programs will have to make sense to multiple computers and operating systems. Software will no doubt be written in a language (like Java) that can translate commands into readable instructions for whatever device needs them. In our scenario, the news collection was made possible by programs running on the various news servers.



JARED SCHNEIDMAN DESIGN

LITHIUM POWER SOURCE, often called a chewing gum cell, is one of the most promising recent developments in mobile power sources. The battery is flat and flexible, like a stick of chewing gum (one of its manufacturers refers to its product as a film battery because its batteries are also reminiscent of film frames). These batteries, which could soon be as thin as 0.2 millimeter, can be manufactured in long, continuous strips, which should reduce production costs. Both NiCd and NiMH cells can also be produced using the chewing gum format.

Big Challenge: Juice

Perhaps the greatest challenge to the creation of the PN, as it is for the electric car or even for your portable compact-disc player, is also one of the most seemingly mundane ones: building better batteries. Long-lasting and flexible batteries will allow the devices in the PN to continue to function without a constant incoming energy supply. Systems known as uninterruptible power supplies will also protect critical components of the PN in case of a power loss. Backup batteries will provide power for the minutes necessary to save all data and shut the system down cleanly. When power is again available, the network's components would be cued to restart automatically.

Among the most significant improvements in batteries are the use of creative packaging techniques to enhance energy density, along with the introduction of lithium ion technology. Lithium offers higher energy densities than earlier rechargeable technologies—nickel cadmium (NiCd) and nickel metal hydride (NiMH) [*see illustration above*].

As long as technology continues to evolve along current capability and cost curves, the personal network should exist at a price comparable to that of today's high-end PCs. Although we see no insurmountable barriers, two major nontechnical stumbling blocks exist. First, users will need to believe that the benefits of the personal network exceed the potential for failures, which would disrupt usage, and for misuse or exposure of personal information. The second is the willingness of hardware, software and network vendors to cooperate in defining and implementing the standards necessary for the seamless integration we envision.

Should these problems be overcome, the worldwide interplay of personal networks could create a global network, which would bring all of us closer together. The so-called six degrees of separation, which says that every person on the earth is only six acquaintances away from any other person, may still rule. But the time and effort required to traverse those six degrees would vanish. Such a world could be a very interesting place indeed. SA

The Authors

BRAD FRIEDLANDER and MARTYN ROETTER are both principals in the management and technology consulting firm Arthur D. Little, based in Cambridge, Mass. Friedlander helps the firm's customers worldwide make effective use of information technology. Roetter currently focuses on assisting network operators and their equipment suppliers as they apply new broadband transmission and switching technologies and network architectures to deploy new services to their customers.

Further Reading

BEING DIGITAL. Nicholas Negroponte. Alfred A. Knopf, 1995.
COMPUTERS. Egil Juliussen in *IEEE Spectrum*, Vol. 34, No. 1, pages 49–54; January 1997.
THE PC OF THE FUTURE. Special issue of *PC Magazine*, Vol. 16, No. 6; March 25, 1997.
WHAT WILL BE: HOW THE NEW WORLD OF INFORMATION WILL CHANGE OUR LIVES. Michael Dertouzos. HarperSanFrancisco (HarperCollins), 1997.

Copyright 1997 Scientific American, Inc.

FAST FACTS ABOUT THE TRANSISTOR

This year more than HALF A BILLION transistors will be manufactured EVERY SECOND.

A typical integrated circuit used in a computer has 3-5 MILLION TRANSISTORS.

During the 1950s, the cost of a transistor went from $45 to $2. Today's transistors, like the millions used in a microchip, COST LESS than 1/100 THOUSANDTH OF A CENT EACH.

If a digital CELL-PHONE was made with vacuum tubes instead of transistors it would occupy a building LARGER THAN THE WASHINGTON MONUMENT.

By the turn of the century, microcircuits will routinely contain ONE BILLION transistors per chip the size of a fingernail. The patterns etched onto these chips will be as complicated as a ROAD MAP OF THE ENTIRE PLANET—SHRUNK TO LILLIPUTIAN SIZE.

LUCENT TECHNOLOGIES; ILLUSTRATIONS BY DUŠAN PETRIČIĆ

Copyright 1997 Scientific American, Inc.